Thompson Sampling-like Algorithms for Stochastic Rising Bandits

Marco Fiandri
Politecnico di Milano

Alberto Maria Metelli
Politecnico di Milano

Francesco Trovò
Politecnico di Milano

Abstract

Stochastic rising rested bandit (SRRB) is a setting where the arms' expected rewards increase as they are pulled. It models scenarios in which the performances of the options grow as a result of an underlying learning process (e.g., online model selection). Even if the bandit literature provides specifically crafted algorithms based on upper-confidence bounds for such a setting, no study about *Thompson sampling* (TS)-like algorithms has been performed so far. The strong regularity of the expected rewards in the SRRB setting suggests that specific instances may be tackled effectively using adapted and sliding-window TS approaches. This work provides novel regret analyses for such algorithms in SRRBs, highlighting the challenges and providing new technical tools of independent interest. Our results allow us to identify under which assumptions TS-like algorithms succeed in achieving sublinear regret and which properties of the environment govern the complexity of the regret minimization problem when approached with TS. Furthermore, we provide a regret lower bound based on a complexity index we introduce. Finally, we conduct numerical simulations comparing TS-like algorithms with state-of-the-art approaches for SRRBs in synthetic and real-world settings.

1 INTRODUCTION

Traditional *multi-armed bandits* (MABs) models Bubeck et al. (2012); Lattimore and Szepesvári (2020) consider *static* environments, assuming arms with expected rewards that do not change during the learning process. However, many real-world applications present a more challenging scenario where the rewards associated with each arm *dynamically* evolve depending on time and/or on the number of times an arm has been played. In particular, two scenarios have been analyzed in the literature: *restless* and *rested* bandits. While the restless MAB setting assumes that the arms' expected reward changes as an effect of nature, in the *rested* MAB setting (Tekin and Liu, 2012), the arms' evolution is triggered by their pull. A large variety of *restless* settings has been analyzed in the past under the name of *non-stationary* bandits, the most studied being the abruptly changing (Garivier and Moulines, 2011) and smoothly changing (Trovò et al., 2020). In contrast, the *rested* setting has raised the attention of the bandit community more recently (Heidari et al., 2016; Seznec et al., 2019; Metelli et al., 2022).

This paper focuses on the *stochastic rising rested bandits* (SRRBs, Metelli et al., 2022) which reflect situations where the arms' expected rewards *increase*, modeling growing trends. An example is represented by the so-called *combined algorithm selection and hyperparameter optimization* (CASH, Thornton et al., 2013; Li et al., 2020), whose goal is to identify the best learning algorithm and the best hyperparameter configuration for a given machine learning task, representing one of the most fundamental problems in *automatic machine learning* (AutoML, Waring et al., 2020). As a further regularity condition, Heidari et al. (2016) and Metelli et al. (2022) consider *concave* evolution of the expected rewards that are suitable to model satiation effects in recommendations (Clerici et al., 2023; Xu et al., 2023). Problems with similar characteristics arise when selecting an algorithm from a predefined set to optimize a given stochastic function, i.e., *online model selection* (Metelli et al., 2022).

The seminal work by Metelli et al. (2022) approached the regret minimization problem in SRRBs by designing a sliding-window algorithm based on upper-confidence bounds able to provide a worst-case regret of the order of $\widetilde{O}(T^{\frac{2}{3}} + \Upsilon(T))$, where T is the learn-

---

Proceedings of the 29$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

ing horizon and $\Upsilon(T)$ a problem-dependent quantity which characterizes the growth rate of the arm expected rewards.[1] However, to the best of our knowledge, there has been no analysis to investigate whether *Thompson sampling*-like algorithms (TS, Kaufmann et al., 2012; Agrawal and Goyal, 2017) can provide good regret guarantees in the SRRB scenario. The interest in studying TS-like algorithms lies in the fact that they are the most widely used bandit algorithms in real-world applications, both easy to implement and with good empirical performance (Scott, 2010; Chapelle and Li, 2011). This paper aims to address the following three questions: (*i*) What modifications should be introduced to the TS algorithms to tackle SRRBs (*ii*) Which properties of the environment govern their performance? (*iii*) What assumptions are necessary to ensure the no-regret property?

Original Contributions. The contributions of the paper are summarized as follows.

- In Section 3, we redefine the notion of suboptimality gap $\overline{\Delta}(n, n')$ (Equation 2) for the SRRB scenario. Based on it, we introduce a novel complexity index $\sigma_{\boldsymbol{\mu}}(T)$ (Equation 3) representing the number of pulls of the optimal arm needed to allow the TS-like algorithm to detect its optimality.
- In Section 4, we present our TS-like algorithms, with a sliding window of size τ, starting with an optional forced exploration phase in which each arm is pulled Γ times. They are instanced with Beta priors, `ET-Beta-SWTS` (Explore Then-Beta-Sliding Window Thompson Sampling), and Gaussian priors, γ-`ET-SWGTS` (Explore Then-Gaussian Sliding Window Thompson Sampling, both illustrated in , Algorithm 1).
- In Section 5, we introduce novel technical tools to conduct the *frequentist* regret analysis of TS-like algorithms that can be of independent interest. Specifically, we generalize the decomposition of the expected number of pulls provided by Theorem 36.2 of Lattimore and Szepesvári (2020) to *general dynamic* environments and sliding-window TS-like algorithms (Lemma 5.1) and we tackle the problem of the underestimation of the optimal arm in the presence of Bernoulli rewards, exploiting log-convexity and stochastic dominance (Lemma 5.2).
- In Section 6, we provide frequentist regret bounds for our TS-like algorithms with no sliding window (i.e., $\tau = T$) for both Beta and Gaussian priors (Theorems 6.2 and 6.3) depending on our novel index of complexity $\sigma_{\boldsymbol{\mu}}(T)$ and on the total variation divergence between specifically defined distributions characterizing the instances. Our bounds are tight for the case of stationary bandits, and, differently from Agrawal and Goyal (2017), they apply to distributions with expected rewards that are not constrained between $[0, 1]$. We analyze the sliding-window approaches (i.e., $\tau < T$) providing analogous frequentist regret bounds in Appendix 7. Additionally, we conceive a lower bound on the regret dependent on the complexity index $\sigma_{\boldsymbol{\mu}}(T)$ we have introduced, and we compare it with the upper bounds showing the tightness on this term.
- In Section 8, we provide numerical simulations to compare the performances of the proposed algorithms with the ones designed for the SRRB setting and restless bandits in terms of cumulative regret on both synthetic scenarios and with real-world data.

The proofs of the results are reported in Appendix B.

2 RELATED LITERATURE

We revise the related literature with a focus on TS algorithms, restless, rising, and rotting bandits.

TS Algorithms. *Thompson sampling* has a long history, originally designed as a heuristic for sequential decision-making (Thompson, 1933). Only in the past decade has it been analyzed (Kaufmann et al., 2012; Agrawal and Goyal, 2017), providing finite-time logarithmic *frequentist* regret in the stochastic stationary bandits. Although the original frequentist analyses are rather involved, general analysis approaches are now available with a more concrete understanding of the intimate functioning of TS methods (Lattimore and Szepesvári, 2020; Baudry et al., 2023).

Restless Bandits. Algorithms designed for stationary bandits (including UCB and TS) have often been extended to *non-stationary* (or *restless*) bandits. Two approaches are available: passive and active. The former iteratively discards the information coming from the far past (either with a *sliding window* or a *discounted* estimator). Examples are `DUCB` (Garivier and Moulines, 2011), `Discounted TS` (Raj and Kalyani, 2017; Qi et al., 2023), `SW-UCB` (Garivier and Moulines, 2011), and `SW-TS` (Trovò et al., 2020). Instead, the latter class uses *change-detection* techniques (Basseville et al., 1993) to decide when to discard old samples. This occurs when a sufficiently evident change affects the arms' expected rewards. Among the active approaches, we mention `CUSUM-UCB` (Liu et al., 2018), `REXP3` (Besbes et al., 2014), `GLR-klUCB` (Besson and Kaufmann, 2019), and `BR-MAB` (Re et al., 2021).

Rising Bandits. Regret minimization in deterministic SRRBs was proposed by Heidari et al. (2016) with algorithms that, when the expected rewards are increasing and concave (i.e., rising), suffer sublinear re-

---

[1] With the $\widetilde{O}(\cdot)$ notation we disregard logarithmic factors w.r.t. the learning horizon T.

gret. The stochastic version of SRRB has been studied by (Metelli et al., 2022) where the authors provide regret bounds of order $\widetilde{O}(T^{\frac{2}{3}} + \Upsilon(T))$, where $\Upsilon(T)$ is a problem-dependent quantity which characterizes the growth rate of the arm expected rewards. This bound is attained by optimistic algorithms using a sliding window approach, namely `R-ed-UCB` and `R-less-UCB`. More recently, approaches presented in Metelli et al. (2022) have been extended to the novel graph-triggered setting (Genalti et al., 2024), interpolating between the rested and restless cases. Furthermore, Amichay and Mansour (2025) provide specific algorithms for the case in which the growing trend is linear. Moreover, Fiandri et al. (2024) has provided the first lower bounds depending on the complexity term $\Upsilon(T)$. The SRRB setting has also been studied in a BAI framework by Mussi et al. (2024), where the authors propose the `R-UCBE` and `R-SR` algorithms, a UCB-inspired and a successive elimination approach, respectively, that provide guarantees for the fixed budget case. Similar results have been obtained in Cella et al. (2021) and Takemori et al. (2024) under a more restrictive parametric model for the expected reward functions. A comparison between the guarantees of the algorithms presented in this paper and the results from Metelli et al. (2022) is provided in Appendix A.

Rotting Bandits. Another setting related to SRRB is the rotting bandits (Seznec et al., 2019; Levine et al., 2017), where the expected rewards decrease either over time or depending on the arm pulls. The authors propose specifically crafted algorithms to address rotting bandits, obtaining sublinear regret guarantees. However, as highlighted in Metelli et al. (2022), these algorithms cannot be applied to the SRRB setting.

3 PROBLEM FORMULATION

We consider an SRRB instance with stochastic rewards. Let $K \in \mathbb{N}$ be the number of arms. Every arm $i \in [\![K]\!]$[2] is associated with an expected reward $\mu_i : \mathbb{N} \to \mathbb{R}$, where $\mu_i(n)$ is the expected reward of arm i when pulled for the n-th time with $n \in \mathbb{N}$. We denote an SRRB instance as $\boldsymbol{\mu} = (\mu_i)_{i \in [\![K]\!]}$. In a rising bandit, the expected reward function $\mu_i(n)$ is non-decreasing.[3]

Assumption 3.1 (Non-Decreasing)**.** *For every arm $i \in [\![K]\!]$ and number of pulls $n \in \mathbb{N}$, let $\gamma_i(n) := \mu_i(n+1) - \mu_i(n)$ be the increment function, it holds that: $\gamma_i(n) \geqslant 0$.*

The learning process occurs over $T \in \mathbb{N}$ rounds, where T is the learning horizon. At every round $t \in [\![T]\!]$, the agent pulls an arm $I_t \in [\![K]\!]$ and observes a random reward $X_{I_t,t} \sim \nu_{I_t}(N_{I_t,t})$, where for every arm $i \in [\![K]\!]$, we have that $\nu_i(N_{i,t})$ is a probability distribution depending on the current number of pulls up to round t, i.e., $N_{i,t} := \sum_{l=1}^{t} \mathbb{1}\{I_l = i\}$, whose expected value is given by $\mu_i(N_{i,t})$. We define the *average expected reward* as: $\overline{\mu}_i(t) := \frac{1}{t} \sum_{l=1}^{t} \mu_i(l)$. The optimal policy constantly plays the arm with the maximum average expected reward, i.e., $i^*(T) := \arg\max_{i \in [\![K]\!]} \overline{\mu}_i(T)$ (highlighting the dependence on T), that we assume w.l.o.g. unique (Heidari et al., 2016).

Suboptimality Gaps. We introduce, for every suboptimal arm $i \neq i^*(T)$ and number of pulls $n, n' \in [\![T]\!]$, the following novel notions of suboptimality gaps defined in terms of the expected reward $\Delta_i(n, n')$ and of average expected reward $\overline{\Delta}_i(n, n')$, formally:

$$\Delta_i(n, n') := \max\{0, \mu_{i^*(T)}(n) - \mu_i(n')\}, \tag{1}$$

$$\overline{\Delta}_i(n, n') := \max\{0, \overline{\mu}_{i^*(T)}(n) - \overline{\mu}_i(n')\}. \tag{2}$$

Notice that, for specific rounds, the gaps $\overline{\Delta}_i(n, n')$ may be negative, however, at the end of the learning horizon T, we are guaranteed that $\overline{\mu}_{i^*(T)}(T) > \overline{\mu}_i(n')$ for all $n' \in [\![T]\!]$. This allows defining, for every suboptimal arm $i \neq i^*(T)$, the minimum number of pulls $\sigma_i(T)$ of the optimal arm needed so that the average expected reward suboptimality gap of arm i is positive and its maximum $\sigma_{\boldsymbol{\mu}}(T)$ over the suboptimal arms, formally:

$$\sigma_i(T) := \min\{l \in [\![T]\!] : \overline{\mu}_{i^*(T)}(l) > \overline{\mu}_i(T)\}, \tag{3}$$

$$\sigma_{\boldsymbol{\mu}}(T) := \max_{i \neq i^*(T)} \sigma_i(T). \tag{4}$$

After $\sigma_{\boldsymbol{\mu}}(T)$ pulls of $i^*(T)$, we can identify $i^*(T)$ as the optimal arm, treating the problem as a stationary bandit since no suboptimal arm can exceed the performance of $i^*(T)$ beyond this point.

Regret. Given an SRRB instance $\boldsymbol{\mu}$, the goal of a regret minimization algorithm $\mathfrak{A}$ is to minimize the *expected cumulative regret* (Heidari et al., 2016; Metelli et al., 2022) defined as follows:

$$R_{\boldsymbol{\mu}}(\mathfrak{A}, T) := T\, \overline{\mu}_{i^*(T)}(T) - \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} \mu_{I_t}(N_{I_t,t})\right], \tag{5}$$

where the expectation is w.r.t. the randomness of the rewards and the possible randomness of $\mathfrak{A}$.

Reward Distribution. We perform the regret analysis under the following reward distributions.

Assumption 3.2 (Bernoulli Rewards)**.** *For every $t \in [\![T]\!]$, $n \in [\![T]\!]$ and $i \in [\![K]\!]$, the reward $X_{i,t} \sim \nu_i(n)$ is Bernoulli distributed.*

Assumption 3.3 (Non-negative Subgaussian re-

---

[2] For two integers $a, b \in \mathbb{N}, a < b$, we denote with $[\![a]\!]$ the set $\{1, \ldots, a\}$ and $[\![a, b]\!]$ the set $\{a, \ldots, b\}$.

[3] Differently from previous works (Cella et al., 2021; Metelli et al., 2022; Mussi et al., 2024), we will not enforce *concavity* since our algorithms will enjoy regret guarantees depending on a novel complexity index (see Equation 3), allowing for sublinear instance-dependent regret guarantees even for non-concave expected reward functions.

wards[4]). *For every $t \in \llbracket T \rrbracket$, $n \in \llbracket T \rrbracket$, and $i \in \llbracket K \rrbracket$, the reward $X_{i,t} \sim \nu_i(n)$ is: (i) non-negative almost surely; (ii) λ^2-subgaussian.*[5]

4 ALGORITHMS

In this section, we introduce `ET-Beta-SWTS` (Explore Then-Beta-Sliding Window Thompson Sampling) and γ-`ET-SWGTS` (γ-Explore Then-Sliding Window Gaussian Thompson Sampling, both illustrated in Algorithm 1), TS-like algorithms designed to deal with Bernoulli (Assumption 3.2) and subgaussian non-negative (Assumption 3.3) rewards, respectively. For the sake of generality, the algorithms are presented in their *sliding window* versions, with a window of size τ, since the non-windowed versions can be retrieved by setting $\tau = T$.

In the first $K\Gamma$ rounds, both algorithms perform a *forced exploration*, playing each arm Γ times, collecting $S_{i,K\Gamma+1,\tau}$ and $N_{i,K\Gamma+1,\tau}$, where $N_{i,t,\tau} := \sum_{s=\max\{t-\tau,1\}}^{t-1} \mathbb{1}\{I_s = i\}$, is the number of times arm $i \in \llbracket K \rrbracket$ has been selected in the last $\min\{t, \tau\}$ rounds with $t \in \llbracket T \rrbracket$, and $S_{i,t,\tau} := \sum_{s=\max\{t-\tau,1\}}^{t-1} X_{i,s} \mathbb{1}\{I_s = i\}$ is the cumulative reward collected by arm i in the last $\min\{t, \tau\}$ rounds. Then, in the subsequent rounds, both algorithms run a TS-like routine. For `ET-Beta-SWTS`, in each round $t \in \llbracket K\Gamma + 1, T \rrbracket$, the posterior distribution of the expected reward of arm i at round t is a *Beta distribution* defined as $\eta_{i,t} := \mathrm{Beta}(S_{i,t,\tau} + 1, N_{i,t,\tau} - S_{i,t,\tau} + 1)$, where $\mathrm{Beta}(\alpha, \beta)$ denotes a Beta distribution with parameters $\alpha > 0$ and $\beta > 0$. Then, for each arm $i \in \llbracket K \rrbracket$, the algorithm draws a sample $\theta_{i,t,\tau}$ from $\eta_{i,t}$, a.k.a. Thompson sample, and plays the arm whose sample has the highest sample value. Then, the values for $S_{i,t+1,\tau}$ and $N_{i,t+1,\tau}$ and the posterior distribution $\eta_{i,t+1}$ are updated. Instead, in γ-`ET-SWGTS`, after the forced exploration, at every round $t \in \llbracket K\Gamma + 1, T \rrbracket$, the posterior distribution is a *Gaussian distribution* $\eta_{i,t} := \mathcal{N}\left(\frac{S_{i,t,\tau}}{N_{i,t,\tau}}, \frac{1}{\gamma N_{i,t,\tau}}\right)$, where $\mathcal{N}(m, s^2)$ is a Gaussian distribution with mean m and variance s^2. The parameter $\gamma > 0$, whose value will be specified later, scales the variance. If there exists an arm $i \in \llbracket K \rrbracket$ from which we do not have samples, i.e., $N_{i,t,\tau} = 0$, the algorithm pulls it, so that the posterior distribution is well defined. Otherwise, the algorithm draws a random sample $\theta_{i,t,\tau}$ from $\eta_{i,t}$ and plays the arm whose Thompson sample is the largest. Finally, the algorithm updates $S_{i,t+1,\tau}$, $N_{i,t+1,\tau}$, and $\eta_{i,t+1}$.

Algorithm 1 `ET-Beta-SWTS` / γ-`ET-SWGTS`.

```
Input: Number of arms $K$, time horizon $T$, time window $\tau$, forced exploration $\Gamma$
Pull each arm $\Gamma$ times
Set $\eta_{i,K\Gamma+1}$ for each $i \in \llbracket K \rrbracket$
for $t \in \llbracket K\Gamma + 1, T \rrbracket$ do
    if $\exists i \in \llbracket K \rrbracket$ such that $N_{i,t,\tau} = 0$ then
        Select $I_t$ as any arm $i$ such that $N_{i,t,\tau} = 0$
    else
        Sample $\theta_{i,t,\tau} \sim \eta_{i,t}$ for each $i \in \llbracket K \rrbracket$
        Select $I_t \in \arg\max_{i \in \llbracket K \rrbracket} \theta_{i,t,\tau}$
    end if
    Pull arm $I_t$ and collect reward $X_{I_t,t}$
    Update $S_{i,t+1,\tau}$ and $N_{i,t+1,\tau}$ for each $i \in \llbracket K \rrbracket$
    Update $\eta_{i,t+1}$ for each $i \in \llbracket K \rrbracket$
end for
```

Note that by removing the forced exploration, i.e., setting $\Gamma = 0$, leaving so the exploration needed for the parameters to be always well defined (represented by if-statement) in `ET-Beta-SWTS`, the algorithm is (almost) identical to `Beta-SWTS` introduced by Trovò et al. (2020); while by setting $\Gamma = 0$ in γ-`ET-SWGTS`, we retrieve a new algorithm, that is a natural extension of `Beta-SWTS` in settings with subgaussian rewards, that we call γ-`SWGTS`.[6]

5 CHALLENGES AND NEW TECHNICAL TOOLS

The scenario where the expected rewards vary throughout the learning process presents additional technical challenges and necessitates the development of new theoretical tools. In this section, we introduce them, highlighting their generality beyond the scope of this paper. To appreciate the content of this section, the interested reader is advised to have a general understanding of the *frequentist* TS analysis of Agrawal and Goyal (2017). The reader not interested in these technical aspects can freely skip this section and proceed to Section 6, which presents the regret bounds.

Let us start by introducing the probability that a Thompson sample is larger than a given threshold.

Definition 5.1. *Let $i, i' \in \llbracket K \rrbracket$ be two arms, $t \in \llbracket T \rrbracket$ be a round, $\tau \in \llbracket T \rrbracket$ be the window size, and $y_{i',t}$ be an arbitrary deterministic threshold, we define: $p_{i,t,\tau}^{i'} := \mathbb{P}\left(\theta_{i,t,\tau} > y_{i',t} \middle| \mathcal{F}_{t-1}\right)$, where $\mathcal{F}_t$ is the filtration induced by the sequence of arms played and observed rewards up to round t.*

The TS analysis of Agrawal and Goyal (2017) is based on bounding the expected number of pulls $\mathbb{E}_{\boldsymbol{\mu}}[N_{i,T}]$

---

[4] For the sake of presentation, we assume positive support. W.l.o.g. one might also consider realizations bounded from below by a given constant (see Section 6 for details).

[5] A zero-mean random variable X is λ^2-subgaussian if for every $s \in \mathbb{R}$ it holds that $\mathbb{E}[e^{sX}] \leqslant e^{s^2\lambda^2/2}$.

[6] In Appendix F, we show that the sliding window approach we used does not increase the per-round computational complexity that remains $O(K)$ as in standard TS with no window.

of every arm $i \in [\![K]\!]$ through a decomposition that was formalized in Theorem 36.2 of Lattimore and Szepesvári (2020) for stationary bandits. We now generalize it for a generic dynamic setting (including restless and rested) and for sliding window approaches. To this end, let $i^*(t) \in [\![K]\!]$ be the optimal arm at time t, $N'_{i,T} := \sum_{t=1}^{T} \mathbb{1}\{I_t = i,\ i \neq i^*(t)\}$ and $\mathcal{T}_i := \{t \in [\![K\Gamma + 1, T]\!]\ :\ i \neq i^*(t)\}$.[7] For the `ET-Beta-SWTS` algorithm, the following holds.

Lemma 5.1 (Expected Number of Pulls Bound for `ET-Beta-SWTS`)**.** *Let $T \in \mathbb{N}$ be the learning horizon, $\tau \in [\![T]\!]$ the window size, $\Gamma \in [\![0, T]\!]$ the forced exploration parameter, for the `ET-Beta-SWTS` algorithm it holds for every $i \in [\![K]\!]$ and free parameter $\omega \in [\![0, T]\!]$ that:*

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mu}}[N'_{i,T}] &\leqslant 1 + \Gamma + \frac{\omega T}{\tau} + \frac{T}{\tau} \\ &+ \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t\in\mathcal{T}_i} \mathbb{1}\left\{p^i_{i,t,\tau} > \frac{1}{T},\ I_t = i,\ N_{i,t,\tau} \geqslant \omega\right\}\right] \\ &+ \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t\in\mathcal{T}_i}\left(\frac{1}{p^i_{i^*(t),t,\tau}} - 1\right)\mathbb{1}\left\{I_t = i^*(t)\right\}\right].\end{aligned}$$

In Appendix B.2, we present an analogous result for γ-`ET-SWGTS` (Lemma B.1). First, we observe that by setting $\omega = \Gamma = 0$ and $\tau = T$, we retrieve Theorem 36.2 of Lattimore and Szepesvári (2020). As highlighted by Agrawal and Goyal (2012); Kaufmann et al. (2012); Agrawal and Goyal (2017); Jin et al. (2023, 2022), the main difficulty in the frequentist analysis of TS-like algorithms lies in the evaluation of the error caused by the underestimation of the optimal arm, measured by the term $\mathbb{E}_{\boldsymbol{\mu}}[1/p^i_{i^*(t),t,\tau}]$. More recently, Baudry et al. (2023) proved that controlling this term is essential to designing optimal TS-like algorithms. Lemma 5.1 confirms that this is true even in dynamic settings, including the restless and rested ones. An additional term Γ arises from the forced exploration, and $\omega > 0$ is a free parameter whose value can be chosen to tighten the bound when a sliding window size τ is used. In what follows, we omit the dependence on τ. The central challenge when bounding the underestimation term $\mathbb{E}_{\boldsymbol{\mu}}[1/p^i_{i^*(t),t}]$ lies in characterizing the *distribution of the parameters of the posterior distribution* of the optimal arm $i^*(t)$, which, in turn, depend on the cumulative reward $S_{i^*(t),t}$. In a stationary bandit, $S_{i^*(t),t}$ is a sum of $N_{i^*(t),t} = j$ *identically distributed* rewards, but, in a dynamic scenario, rewards are no longer identically distributed. While this does not pose significant issues for Gaussian priors, it prevents us from applying the technique of Agrawal and Goyal (2017) for Beta priors and Bernoulli rewards. Indeed, such an analysis heavily relies on the fact that $S_{i^*(t),t}$ (i.e., the number of successes) is a *Binomial distribution*, being the sum of identically distributed Bernoulli distributions with parameter $\overline{\mu}_{i^*(t),j}$. Instead, in the dynamic setting, $S_{i^*(t),t}$ is a sum of non-identically distributed Bernoulli distributions (with parameters $\mu_{i^*(t),1}, \dots, \mu_{i^*(t),j}$) since the arm expected reward changes throughout the learning process. Instead, it can be proved that $S_{i^*(t),t}$ is distributed as a *Poisson-Binomial* (Wang, 1993), whose analytical treatment is far more challenging than the binomial counterpart. Let us first instance Definition 5.1 for Beta priors: $p^i_{i^*(t),t} = \mathbb{P}(\mathrm{Beta}(S_{i^*(t),t} + 1, F_{i^*(t),t} + 1) > y_{i,t} | \mathcal{F}_{t-1})$, where $F_{i^*(t),t} = N_{i^*(t),t} - S_{i^*(t),t}$ (i.e., the number of failures). Our technical innovation consists in Lemma 5.2, presented below, in which we show that, surprisingly, the worst case for the underestimation is attained when all rewards are identically distributed, i.e., when $S_{i^*(t),t}$ has a Binomial distribution.

Lemma 5.2 (PB-Bin Stochastic Dominance)**.** *Let $j \in \mathbb{N}$, $\mathrm{PB}(\underline{\mu}_{i^*(t)}(j))$ be a Poisson-Binomial distribution with parameters $\underline{\mu}_{i^*(t)}(j) = (\mu_{i^*(t),1}, \dots, \mu_{i^*(t),j})$, and $\mathrm{Bin}(j, x)$ be a binomial distribution of j trials and success probability $0 \leqslant x \leqslant \frac{1}{j}\sum_{l=1}^{j} \mu_{i^*(t),l} = \overline{\mu}_{i^*(t),j}$. Then, it holds that:*

$$\begin{aligned}&\mathop{\mathbb{E}}_{S_{i^*(t),t} \sim \mathrm{PB}(\underline{\mu}_{i^*(t)}(j))}\left[\frac{1}{p^i_{i^*(t),t}}\middle| N_{i^*(t),t} = j\right] \\ &\qquad \leqslant \mathop{\mathbb{E}}_{S_{i^*(t),t} \sim \mathrm{Bin}(j,x)}\left[\frac{1}{p^i_{i^*(t),t}}\middle| N_{i^*(t),t} = j\right].\end{aligned}$$

The result is derived by: (i) showing the *discrete log-convexity* of $1/p^i_{i^*(t),t}$ as a function of the number of successes (Hoggar, 1974; Johnson and Goldschmidt, 2006; Hill and Houdré, 1999) to pass from the expectation w.r.t. the Poisson-Binomial to that w.r.t. the binomial and obtain the first inequality and (ii) relying on *stochastic dominance* (Boland et al., 2002, 2004; Marshall et al., 2011) arguments to show the monotonicity of the expectation w.r.t. the parameter of the binomial distribution and obtain the second inequality. A similar result is proven for Gaussian priors, resorting, however, to more standard arguments based on concentration inequalities (i.e., Chernoff bounds). Lemma 5.2 applies to: (i) *any dynamic environment* in which the optimal arm $i^*(t)$ may change across rounds and (ii) *sliding window* approaches. For these reasons, it represents our main technical novelty and can be leveraged for the analysis of TS-like algorithms in dynamic settings even beyond the scope of the paper.

6 REGRET ANALYSIS

We start this section by providing a general regret analysis of `ET-Beta-TS` and γ-`ET-GTS` in the SRRBs

---

[7] For the SRRB setting, $i^*(t) = i^*(T)$ for every $t \in [\![T]\!]$ and $N'_{i,T} = N_{i,T}$ for $i \in [\![K]\!] \backslash \{i^*(T)\}$.

setting, i.e., the algorithms presented in Section 4 with sliding window $\tau = T$ (Section 6.1). The regret bounds of the corresponding windowed versions, i.e., `ET-BETA-SWTS` and γ-`ET-SWGTS`, are reported in Appendix 7. We conclude the section by deriving a regret lower bound and discussing its tightness (Section 6.2).

6.1 Regret Upper Bounds of ET-TS Algorithms for SRRBs

Before presenting the results, we introduce the following lemma relating the expected cumulative regret to the expected number of pulls of the suboptimal arms.

Lemma 6.1 (Wald's Inequality for Rising Bandits)**.** *Let* $\mathfrak{A}$ *be an algorithm and* $T \in \mathbb{N}$ *be a learning horizon, it holds that:* $R_{\boldsymbol{\mu}}(\mathfrak{A}, T) \leqslant \sum_{i \neq i^*(T)} \Delta_i(T,1)\, \mathbb{E}_{\boldsymbol{\mu}}[N_{i,T}]$.

Note that Lemma (6.1) holds with equality for stationary bandits. In the following, as it is typical in the literature (e.g., Garivier and Moulines, 2011; Besson and Kaufmann, 2019; Trovò et al., 2020), we provide bounds on the expected number of pulls of the suboptimal arms, under different assumptions for the reward distributions (Assumptions 3.2 and 3.3). From these results, the regret bound can be easily computed via Lemma 6.1. We start with the `ET-Beta-TS` algorithm.

Theorem 6.2 (`ET-Beta-TS` Bound)**.** *Let* $\sigma \in [\![\sigma_{\boldsymbol{\mu}}(T), T]\!]$*. Under Assumptions 3.1 and 3.2, for the* `ET-Beta-TS` *algorithm, for every arm* $i \in [\![K]\!] \backslash \{i^*(T)\}$*, it holds for every* $\epsilon \in (0,1]$*, that:*[8]

$$\mathbb{E}_{\boldsymbol{\mu}}[N_{i,T}] \leqslant O\Bigg(\underbrace{\Gamma}_{(i)} + \underbrace{\frac{(1+\epsilon)\log(T)}{d(\overline{\mu}_i(T), \overline{\mu}_{i^*(T)}(\sigma))} + \frac{1}{\epsilon^2}}_{(ii)} + \underbrace{\sum_{j=\Gamma}^{\sigma} \frac{\delta_{\mathrm{TV}}(\mathrm{Bin}(j, \overline{\mu}_{i^*(T)}(j)), \mathrm{Bin}(j, \overline{\mu}_{i^*(T)}(\sigma)))}{(1-\overline{\mu}_{i^*(T)}(\sigma))^{j+1}}}_{(iii)} \Bigg),$$

where $d(x,y) := x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y}$ *is the KL divergence between Bernoulli distributions of parameters* $x, y \in [0,1]$*,* $\delta_{\mathrm{TV}}(P,Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$ *is the total variation between* P *and* Q*,* $\mathrm{Bin}(j,x)$ *is the binomial distribution with* j *trials and parameter* x*.*

Let us now move to the γ-`ET-GTS` algorithm.

Theorem 6.3 (γ-`ET-GTS` Bound)**.** *Let* $\sigma \in [\![\sigma_{\boldsymbol{\mu}}(T), T]\!]$*. Under Assumptions 3.1 and 3.3, setting* $\gamma \leqslant \min\left\{\frac{1}{4\lambda^2}, 1\right\}$*, for the* γ-`ET-GTS` *algorithm, for every arm* $i \in [\![K]\!] \backslash \{i^*(T)\}$*, it holds that:*

$$\mathbb{E}_{\boldsymbol{\mu}}[N_{i,T}] \leqslant O\Bigg(\underbrace{\Gamma}_{(i)} + \underbrace{\frac{\log(T\overline{\Delta}_i(\sigma,T)^2 + e^6)}{\gamma\overline{\Delta}_i(\sigma,T)^2}}_{(ii)} + \underbrace{\sum_{j=\Gamma}^{\sigma} \frac{\delta_{\mathrm{TV}}(\mathbb{P}_j, \mathbb{Q}_j(\overline{\mu}_{i^*(T)}(\sigma)))}{\mathrm{erfc}(\sqrt{\gamma j/2}(\overline{\mu}_{i^*(T)}(\sigma)))}}_{(iii)} \Bigg),$$

where $\mathrm{erfc}(\cdot)$ *is the complementary error function,* $\mathbb{P}_j$ *is the distribution of the sample mean of the first* j *samples collected from arm* $i^*(T)$*, while* $\mathbb{Q}_j(y)$ *is the distribution of the sample mean of* j *samples collected from* any λ^2*-subgaussian distribution with mean* y*.*

Both bounds on the expected number of pulls contain three terms: (i) is the number of pulls Γ during the *forced exploration*; (ii) is the *expected number of pulls in a stationary bandit* with expected rewards $\overline{\mu}_i(T)$ and $\overline{\mu}_{i^*(T)}(\sigma)$; (iii) is cumulative *total variation (TV) distance* measuring the dissimilarity between the real distribution of the optimal arm's rewards and that of a stationary bandit with expected reward $\overline{\mu}_{i^*(T)}(\sigma)$ which originates from a *change of measure* argument.[9] Focusing on term (iii), for `ET-Beta-TS`, it is related to the TV between the distribution of the reward at the j-th pull $\mathrm{Bin}(j, \overline{\mu}_{i^*(T)}(j))$ and that evaluated the reference number of pulls $\mathrm{Bin}(j, \overline{\mu}_{i^*(T)}(\sigma))$. Similarly, for γ-`ET-GTS`, (iii) is related to the TV divergence between the distribution $\mathbb{P}_j$ of the sample mean of the first j rewards sampled from the optimal arm $i^*(T)$ and the distribution $\mathbb{Q}_j$ of the sample mean of j samples obtained from *an arbitrarily chosen* subgaussian distribution with mean $\overline{\mu}_{i^*(T)}(\sigma)$ and parameter λ^2. These terms vanish in a stationary bandit since, by setting $\Gamma = 0$, we retrieve the state-of-the-art bounds for TS with Bernoulli and Gaussian priors (see Theorems 1.1 and 1.3 from Agrawal and Goyal (2017)). The free parameter $\sigma \in [\![\sigma_{\boldsymbol{\mu}}(T), T]\!]$ can be chosen to tighten the bounds, exercising the tradeoff between term (ii), that decreases with σ since $\overline{\mu}_{i^*(T)}(\sigma)$ moves away from $\overline{\mu}_i(T)$ and term (iii), that increases with σ being the summation made of $\sigma - \Gamma$ terms. Whenever σ can be selected as a constant independent from T (i.e., when $\sigma_{\boldsymbol{\mu}}(T)$ is independent of T), term (iii) is a constant of order $O(\sigma)$, leading to a regret bound that matches that of stationary bandits. In these cases, we can freely remove the forced exploration by setting $\Gamma = 0$. However, whenever σ depends on T, the bounds suggest that enforcing the exploration through Γ can be beneficial since term (i) can be smaller than term (iii) when the TV divergence is close to 1 (due to the denominators in the summations of term (iii)).

---

[8] For the sake of presentation, with the Big-O notation, we are neglecting terms depending on $\mu_i(T)$ and $\mu_{i^*(T)}(\sigma)$, but not explicitly on T. The full expression is reported in Equation (112) in the appendix.

[9] The state-of-the-art bounds for the total variation distance are reported in Lemmas C.4 and C.15.

Finally, we note that in γ-`ET-GTS`, the non-negativity of the reward, enforced by Assumption 3.3, is needed to bound the denominator of the addenda in terms (iii).[10] We highlight that Theorems 6.2 and 6.3 hold for every SRRB under the assumption that the expected reward functions are non-decreasing (Assumption 3.1) only, with no need of enforcing the *concavity* assumption. We remark that our bounds are *instance-dependent* (w.r.t. $\overline{\Delta}_i$ and $\sigma_{\boldsymbol{\mu}}$). Note that, as shown in previous works (Metelli et al., 2022; Fiandri et al., 2024), the *worst-case* regret over the class of SRRBs degenerates to linear if no further structure is enforced (concavity is not enough), showing that the problem unlearnable in the minimax sense.

6.2 Explicit Regret Upper and Lower Bounds

In this section, we first make use of Theorems 6.2 and 6.3 to obtain more explicit upper bounds, we derive a regret lower bound, and we discuss the upper bounds' tightness. We introduce a subset of SRRBs parametrized by a bound $\overline{\sigma} \geqslant 0$ to the complexity index $\sigma_{\boldsymbol{\mu}}(T)$: $\mathcal{M}_{\overline{\sigma}} := \{\boldsymbol{\mu} \text{ SRRB } : \sigma_{\boldsymbol{\mu}}(T) \leqslant \overline{\sigma}\}$. We denote with $\mathcal{M}_{\overline{\sigma}}^{\det}$ the subset of the *deterministic* SRRBs in $\mathcal{M}_{\overline{\sigma}}$. Intuitively, $\overline{\sigma}$ controls the complexity of the SRRB instances, as it corresponds to the number of pulls needed to distinguish the optimal arm $i^*(T)$ from the suboptimal ones in the worst instance of $\mathcal{M}_{\overline{\sigma}}$. In particular, if $\overline{\sigma}' \geqslant \overline{\sigma}$, we have that $\mathcal{M}_{\overline{\sigma}} \subseteq \mathcal{M}_{\overline{\sigma}'}$, and $\mathcal{M}_T$ is the set of all SRRBs. This allows deriving a more explicit regret bound for our ET-TS algorithms.

Corollary 6.4 (Explicit `Beta-TS` and γ-`GTS` Bound)**.** *Let $\overline{\sigma} \geqslant 0$. Under the same assumptions of Theorems 6.2 and 6.3, setting $\Gamma = \alpha\overline{\sigma}$, with $\alpha \geqslant 1$, for both `ET-Beta-TS` and γ-`ET-GTS`, for every $\boldsymbol{\mu} \in \mathcal{M}_{\overline{\sigma}}$ and for every arm $i \in [\![K]\!] \backslash \{i^*(T)\}$, it holds that:*[11]

$$\mathbb{E}_{\boldsymbol{\mu}}[N_{i,T}] \leqslant O\Bigg(\underbrace{\overline{\sigma}}_{(i)} + \underbrace{\frac{\log(T)}{\overline{\Delta}_i(\overline{\sigma}, T)^2}}_{(ii)} \Bigg).$$

Thus, we identify two components: (i) the complexity index $\overline{\sigma}$ and (ii) the expected number of pulls $O\left(\frac{\log(T)}{\overline{\Delta}_i^2}\right)$ unavoidable even in stationary stochastic bandit (Lattimore and Szepesvári, 2020). The following regret lower bound shows that the dependence on $\overline{\sigma}$ is unavoidable even for the deterministic SRRBs.

Theorem 6.5 (Lower Bound)**.** *Let $T \in \mathbb{N}$ and $\overline{\sigma} \in [\![2, \frac{T-1}{2}]\!]$. For every algorithm $\mathfrak{A}$, it holds that:*

$$\sup_{\boldsymbol{\mu} \in \mathcal{M}_{\overline{\sigma}}^{\det}} R_{\boldsymbol{\mu}}(\mathfrak{A}, T) \geqslant \frac{K}{64}(\overline{\sigma} - 2). \tag{6}$$

First of all, we note that the lower bound is derived considering deterministic instances. Indeed, the complexity index $\sigma_{\boldsymbol{\mu}}(T)$ is not affected by the possible reward stochasticity. Thus, *any algorithm* that wishes to be no-regret (with no additional information), cannot avoid playing every arm a number of times proportional to $\overline{\sigma}$, as it is always possible to design an instance that needs at least $\overline{\sigma}$ pulls to differentiate $i^*(T)$ from the suboptimal arms. Finally, by recalling that the logarithmic regret component in Corollary 6.4 comes from the reward stochasticity (and it is tight for stationary bandits), Theorem 6.5 shows that our bound of Corollary 6.4 is tight in the dependence on the index bound $\overline{\sigma}$ for the class of SRRBs $\mathcal{M}_{\overline{\sigma}}$.

7 SLIDING WINDOW APPROACHES

The main drawback of the previously presented approach is that they use *all the samples from the beginning of learning* for estimating the average expected reward. However, in some cases, it might be convenient to forget the past and focus on the most recent samples only.

Preliminaries. We extend the definitions of Section 3 to account for a sliding window. For every arm $i \in [\![K]\!]$, round $t \in [\![T]\!]$, and window size $\tau \in [\![T]\!]$, we define the *windowed average expected reward* as $\overline{\mu}_i(s;\tau) := \frac{1}{\tau}\sum_{l=s-\tau+1}^{s} \mu_i(l)$. Furthermore, we define the minimum number of pulls needed so that the optimal arm $i^*(T)$ can be identified as optimal in a window of size τ:

$$\sigma_i'(T;\tau) := \min\{\{s \in [\![T]\!] : \overline{\mu}_{i^*(T)}(s;\tau) > \mu_i(T)\} \cup \{+\infty\}\},$$
$$\sigma'(T;\tau) := \max_{i \neq i^*(T)} \sigma_i'(T;\tau).$$

These definitions resemble those of Equation (3). However, here they involve the windowed average expected reward of the optimal arm $\overline{\mu}_{i^*(T)}(s;\tau)$ compared against the expected reward (not averaged) of the other arms at the end of the learning horizon $\mu_i(T)$. Furthermore, for some values of τ, a number of pulls s so that $\overline{\mu}_{i^*(T)}(s;\tau) > \mu_i(T)$ might not exist. In such a case, we set $\sigma_i'(T;\tau)$ (and, thus, $\sigma'(T;\tau)$) to $+\infty$. Nevertheless, as visible in Figure 1, in some cases $\sigma'(T;\tau) \ll \sigma(T)$. This justifies the introduction of the complexity index $\sigma'(T;\tau)$ that will appear in the regret bounds presented in this section. Finally, we introduce a new definition of suboptimality gaps: $\Delta_i'(T;\tau) := \overline{\mu}_{i^*(T)}(\sigma'(T;\tau);\tau) - \mu_i(T)$ for every arm $i \in [\![K]\!]$. In the following for the sake of presenta-

---

[10] For rewards $X \geqslant -b$ for some $b \geqslant 0$ a.s., we can replace the denominator in term (iii) of Theorem 6.3 with $\operatorname{erfc}(\sqrt{\gamma j/2}(\overline{\mu}_{i^*(T)}(\sigma) + b))$ (see Equation 132).

[11] Notice that the choice of the exploration parameter Γ requires just an upper bound $\overline{\sigma}$ to the complexity index $\sigma_{\boldsymbol{\mu}}(T)$ of the set of SRRB of interest $\mathcal{M}_{\overline{\sigma}}$. Details are provided in Appendix D.

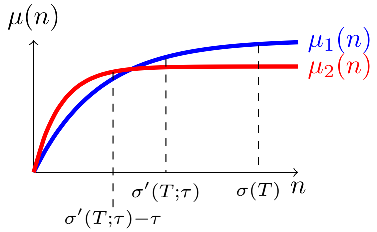


Figure 1: Visual representation of $\sigma'(T;\tau)$, a point in which $\overline{\mu}_{i*(T)}(\sigma'(T;\tau),\tau) > \mu_i(T)$.

tion we will denote $\sigma'(T;\tau)$ as σ', $\Delta_i'(T;\tau)$ as Δ_i' and $\Delta' := \min_{i\in[\![K]\!]\setminus\{i*(T)\}} \Delta_i'$.

Regret. In this section, we analyze the sliding window versions of TS with Bernoulli priors, namely `ET-Beta-SWTS` (Algorithm 1), and Gaussian priors, namely γ-`ET-SWGTS` (Algorithm 1). The following results provide the regret upper bounds achieved by these algorithms as a function τ.

Theorem 7.1 (`ET-Beta-SWTS` Bound)**.** *Under Assumption 3.2, for the* `ET-Beta-SWTS` *algorithm with a window of size $\tau \in [\![T]\!]$, for every arm $i \in [\![K]\!]\setminus\{i^*(T)\}$, it holds that:*

$$\mathbb{E}_{\boldsymbol{\mu}}[N_{i,T}] \leqslant O\Bigg(\underbrace{\Gamma}_{(i)} + \underbrace{\frac{T\log(T)}{\tau\Delta_i'^2}}_{(ii)} + \underbrace{(\sigma'-\Gamma)\tau}_{(iii)}\Bigg).$$

Theorem 7.2 (γ-`ET-SWGTS` Bound)**.** *Under Assumption 3.3, setting $\gamma \leqslant \min\left\{\frac{1}{4\lambda^2}, 1\right\}$, for the* γ-`ET-SWGTS` *algorithm with a window of size $\tau \in [\![T]\!]$, for every arm $i \in [\![K]\!]\setminus\{i^*(T)\}$, it holds that:*

$$\mathbb{E}_{\boldsymbol{\mu}}[N_{i,T}] \leqslant O\Bigg(\underbrace{\Gamma}_{(i)} + \underbrace{\frac{T\log(T\Delta_i'^2+e^6)}{\gamma\tau\Delta_i'^2} + \frac{T}{\tau}}_{(ii)} + \underbrace{(\sigma'-\Gamma)\tau}_{(iii)}\Bigg).$$

We recognize the terms related to (i) the number of pulls Γ performed during the *forced exploration* ; (ii) the *expected number of pulls for a windowed algorithm in standard bandits*; (iii) the sum of exponential terms of Theorem 6.2 and Theorem 6.3 is replaced by a *linear term proportional to σ' and τ*. First, the regret bounds are presented for a generic choice of the window size τ, whose optimal choice depends on the instance-dependent quantities σ' and Δ_i'. Second, if we compare these regret bounds with those of the corresponding non-windowed versions, we observe that the advantage of a sliding window approach is twofold: first, (iii) has a magnitude proportional $O(\sigma) = O(\sigma(T))$ in Theorems 6.2 and 6.3 and becomes $O(\sigma')$ in Theorems 7.1 and 7.2. This quantifies the advantage of the sliding window algorithms in the cases in which $\sigma' \ll \sigma(T)$. Second, the proportionality is not exponential anymore, but linear in the complexity term. Thus, the windowed algorithms can be more robust to inaccurate choices of Γ (i.e., Γ too small). In fact, since we discard samples, we are able to control the bias. Again, the forced exploration can be useful to reduce the impact of the bias term, reducing linearly the dependance on σ'. The dependence on T and τ in (ii) is in line with the state-of-the-art results for windowed algorithms (Combes and Proutiere, 2014, Theorem 5.1). Finally, we remark that these regret bounds become vacuous when $\sigma' = +\infty$. As it is common in the Sliding-Window literature (Trovò et al., 2020; Garivier and Moulines, 2011), the best choice for the window-lenght is instance dependent. However, the analysis highlights that this design offers practitioners a principled way to trade off adaptivity and variance with the choice of the window length, a parameter under their control. Whenever it exists a window length such that $\tau\sqrt{\sigma'} = \Theta(\sqrt{T\log(T)})$[12], for $\Gamma = 0$, considering expected rewards within $[0,1]$, we have for both algorithms:

$$R_{\boldsymbol{\mu}}(\mathfrak{A}, T) \leqslant O\left(\frac{1}{\Delta'^2}\sqrt{\sigma' T\log(T)}\right)$$

8 NUMERICAL SIMULATIONS

8.1 Synthetic Experiments

We consider the same 15-arms environment of Metelli et al. (2022) (Figure 3) and compare against `R-ed-UCB` (Metelli et al., 2022).[13] The parameters, complying with the recommendation of each algorithm, and the parameters defining the environment are provided in Appendix E. We evaluate empirical cumulative regret $\hat{R}(\mathfrak{A}, t)$ averaged over 50 independent runs (with the corresponding standard deviations), over a time horizon of $T = 100,000$ rounds. We run the sliding-window algorithms against the worst-case misspecification for the forced exploration parameter, i.e., $\Gamma = 0$ for both `Beta-SWTS` and γ-`SWGTS`, since the analysis shows that sliding window algorithms are robust to misspecification of the exploration parameter. For the non-windowed algorithms (`ET-Beta-TS` and γ-`ET-GTS`), we consider $\Gamma = 2000$. We also include their standard versions (designed for the stationary setting) without the forced exploration (`Beta-TS` and γ-`GTS`), to assess whether the proposed modifications behave as suggested by our analysis.

Results. In Figure 3, we observe that `Beta-TS`, γ-`SWGTS`, and `Beta-SWTS` suffer smaller regret than `R-ed-UCB` over the entire time horizon, whereas

---

[12] $f(x) = \Theta(g(x))$ if there exist two absolute constants $a, b > 0$ s.t. for all the x in the domain of f and g we have $bg(x) \leqslant f(x) \leqslant ag(x)$

[13] We also compare with the other baselines considered in Metelli et al. (2022) in Appendix E.4.

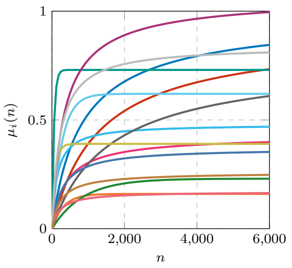


Figure 2: 15-arm setting: expected reward functions of the arms.

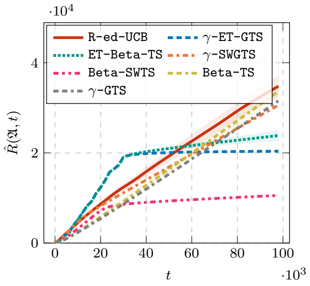
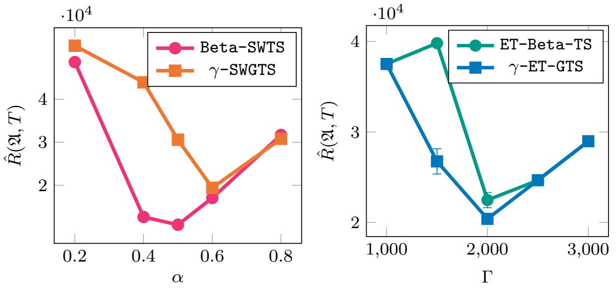


Figure 3: Average cumulative regret in the 15-arm setting (50 runs ± std).



(a) Sensitivity on $\tau = T^{\alpha}$. (b) Sensitivity on Γ.

Figure 4: Average cumulative regret of (a) `Beta-SWTS` and γ-`SWGTS` for different window sizes $\tau = T^{\alpha}$ (b) `ET-Beta-SWTS` and γ-`ET-SWGTS` for different forced exploration Γ.

`ET-Beta-TS` and γ-`ET-GTS`, coherently with the theoretical analysis, suffer linear regret in the forced exploration phase and then catch up to `R-ed-UCB` for $t \gtrsim 50,000$. Indeed, there is evidence of the superiority of our TS-like from that point on. Moreover, all the algorithms, except `Beta-TS`, γ-`GTS` and γ-`SWGTS`, display a flattening regret curve for $t \gtrsim 30,000$. This is explained since `Beta-TS` and γ-`GTS` are designed for stationary bandits, while, as we shall see, the performance of γ-`SWGTS` depends on the window size.

Sensitivity Analysis. We run a sensitivity analysis on the sliding window τ for `Beta-SWTS` and γ-`SWGTS` and on the Γ parameter for `ET-Beta-TS` and γ-`ET-GTS`. We provide the value of their average regret $\hat{R}(\mathfrak{A}, T)$ at the end of the time horizon T over 50 runs on the same 15-arms setting. The results are reported in Figure 4.[14] Coherently with our analysis, whenever the forced exploration Γ or the window size τ is large enough, our algorithms are able to identify the best arm consistently. A too-small value of Γ, instead, leads to a poor exploration that makes the regret grow fast. Similarly, employing a too-small window size τ introduces a large variance, resulting in a poorer performance.

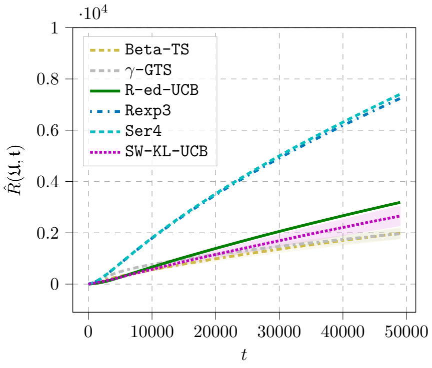


Figure 5: Average cumulative regret in the IMBD setting (30 runs ± std).

8.2 IMDB Experiment

We consider the same *online model selection* task on the IMDB dataset proposed in Metelli et al. (2022), where each arm corresponds to an online-trained classifier and the learner gets a reward 1 when the predicted label is correct, 0 otherwise. Whenever a classifier is pulled, one step of training is performed. We compare the non-windowed approaches (`Beta-TS` and γ-`GTS`), over a time horizon of $T = 50,000$, considering the empirical cumulative regret over 30 runs, against several baselines considered in (Metelli et al., 2022) with the values of the hyperparameters defined there.

Results. The results are reported in Figure 5. We observe that, as we discussed, TS-like algorithms can still be effective whenever the complexity term $\sigma_{\boldsymbol{\mu}}(T)$ is low enough. Indeed, `Beta-TS` and standard γ-`GTS` outperform the baselines, even those designed explicitly for SRRB bandits, i.e., `R-ed-UCB`. Additional experiments are reported in Appendix E.5.

9 CONCLUSIONS

In this paper, we investigated the properties of TS-like algorithms for regret minimization in the setting of SRRBs. We analyzed the `TS` algorithms with Beta and Gaussian priors. In both cases, we derived a general analysis that highlights the challenges of the setting, which makes use of novel technical tools of independent interest. We showed that in the SRRB setting, the classical logarithmic regret is increased by a term that depends on the total variation distance between pairs of suitably defined distributions. We also specialize these results for a parametric subset of SRRBs, showing the tightness of the bounds on the regret, by inferring the complexity of the problem via a lower bound, dependent on a new complexity index. Finally, we provided numerical experiments that demonstrate the advantages of our proposed approaches on both synthetic and real-world data.

---

[14] We report the standard deviation bars that, sometimes, are not clearly visible due to their limited size.

Acknowledgments

This publication was funded with the contribution of Ministero dell'Università e della ricerca pursuant to D.D. n. 7206 of 17 April 2025 – BANDO FIS 2. Project FIS-2023-02598 (Starting Grant), title: "Unified Learning from Diverse Human Feedback" (HUmLrn). CUP: D53C25000710001

References

Abramowitz, M. and Stegun, I. A. (1968). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office.

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory (COLT)*, pages 39–1.

Agrawal, S. and Goyal, N. (2013). Further optimal regret bounds for thompson sampling. In Carvalho, C. M. and Ravikumar, P., editors, *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 31 of *Proceedings of Machine Learning Research*, pages 99–107. PMLR.

Agrawal, S. and Goyal, N. (2017). Near-optimal regret bounds for thompson sampling. *Journal of the ACM*, 64(5):1–24.

Allesiardo, R., Feraud, R., and Maillard, O.-A. (2017). The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics*, 3.

Amichay, O. and Mansour, Y. (2025). Rising rested mab with linear drift. *arXiv preprint arXiv:2501.04403*.

Basseville, M., Nikiforov, I. V., et al. (1993). *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs.

Baudry, D., Suzuki, K., and Honda, J. (2023). A general recipe for the analysis of randomized multi-armed bandit algorithms. *arXiv preprint arXiv:2303.06058*.

Besbes, O., Gur, Y., and Zeevi, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems (NeurIPS)*, volume 27, pages 199–207.

Besson, L. and Kaufmann, E. (2019). The generalized likelihood ratio test meets klucb: an improved algorithm for piece-wise non-stationary bandits. *HAL*, 2019(0).

Boland, P. J., Singh, H., and Cukic, B. (2002). Stochastic orders in partition and random testing of software. *Journal of Applied Probability*, 39(3):555–565.

Boland, P. J., Singh, H., and Cukic, B. (2004). The stochastic precedence ordering with applications in sampling and testing. *Journal of Applied Probability*, 41(1):73–82.

Bubeck, S., Cesa-Bianchi, N., et al. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122.

Cella, L., Pontil, M., and Gentile, C. (2021). Best model identification: A rested bandit formulation. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 1362–1372. PMLR.

Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 24, pages 2249–2257.

Clerici, G., Laforgue, P., and Cesa-Bianchi, N. (2023). Linear bandits with memory: from rotting to rising. *arXiv preprint arXiv:2302.08345*.

Combes, R. and Proutiere, A. (2014). Unimodal bandits: Regret lower bounds and optimal algorithms. In *International Conference on Machine Learning (ICML)*, pages 521–529.

Fiandri, M., Metelli, A. M., and Trovo, F. (2024). Rising rested bandits: Lower bounds and efficient algorithms. *arXiv preprint arXiv:2411.14446*.

Fiandri, M., Metelli, A. M., and Trovò, F. (2025). Sliding-window thompson sampling for non-stationary settings.

Garivier, A. and Cappé, O. (2011). The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Conference On Learning Theory (COLT)*, pages 359–376.

Garivier, A., Lattimore, T., and Kaufmann, E. (2016). On explore-then-commit strategies. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Garivier, A. and Moulines, E. (2011). On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 174–188.

Genalti, G., Mussi, M., Gatti, N., Restelli, M., Castiglioni, M., and Metelli, A. M. (2024). Graph-triggered rising bandits. In *International Conference on Machine Learning (ICML)*. OpenReview.net.

Gil, M., Alajaji, F., and Linder, T. (2013). Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 249:124–131.

Heidari, H., Kearns, M. J., and Roth, A. (2016). Tight policy regret bounds for improving and decaying bandits. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1562–1570.

Hill, T. P. and Houdré, C. (1999). *Advances in Stochastic Inequalities*, volume 234. American Mathematical Society.

Hoggar, S. (1974). Chromatic polynomials and logarithmic concavity. *Journal of Combinatorial Theory, Series B*, 16(3):248–254.

Jin, T., Xu, P., Xiao, X., and Anandkumar, A. (2022). Finite-time regret of thompson sampling algorithms for exponential family multi-armed bandits. In *Advances in Neural Information Processing Systems*, volume 35, pages 38475–38487. Curran Associates, Inc.

Jin, T., Yang, X., Xiao, X., and Xu, P. (2023). Thompson sampling with less exploration is fast and optimal. In *International Conference on Machine Learning (ICML)*, pages 15239–15261. PMLR.

Johnson, O. and Goldschmidt, C. (2006). Preservation of log-concavity on summation. *ESAIM: Probability and Statistics*, 10:206–215.

Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 199–213.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Levine, N., Crammer, K., and Mannor, S. (2017). Rotting bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 3074–3083.

Li, Y., Jiang, J., Gao, J., Shao, Y., Zhang, C., and Cui, B. (2020). Efficient automatic cash via rising bandits. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 4763–4771.

Liu, F., Lee, J., and Shroff, N. (2018). A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 32.

Marshall, A. W., Olkin, I., and Arnold, B. C. (2011). Inequalities: Theory of majorization and its applications. *Springer Series in Statistics*.

Metelli, A. M., Trovò, F., Pirola, M., and Restelli, M. (2022). Stochastic rising bandits. In *International Conference on Machine Learning (ICML)*, pages 15421–15457.

Mussi, M., Montenegro, A., Trovò, F., Restelli, M., and Metelli, A. M. (2024). Best arm identification for stochastic rising bandits. In *International Conference on Machine Learning (ICML)*.

Poisson, S.-D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile: précédées des règles générales du calcul des probabilités*. Bachelier.

Qi, H., Wang, Y., and Zhu, L. (2023). Discounted thompson sampling for non-stationary bandit problems. *arXiv preprint arXiv:2305.10718*.

Raj, V. and Kalyani, S. (2017). Taming non-stationary bandits: A bayesian approach. *arXiv preprint arXiv:1707.09727*.

Re, G., Chiusano, F., Trovò, F., Carrera, D., Boracchi, G., and Restelli, M. (2021). Exploiting history data for nonstationary multi-armed bandit. In *European Conference on Machine Learning (ECML)*, pages 51–66.

Rigollet, P. and Hütter, J.-C. (2023). High-dimensional statistics. *arXiv preprint arXiv:2310.19244*.

Roch, S. (2024). *Modern discrete probability: An essential toolkit*. Cambridge University Press.

Roos, B. (2001). Binomial approximation to the poisson binomial distribution: The krawtchouk expansion. *Theory of Probability & Its Applications*, 45(2):258–272.

Samuels, S. M. (1965). On the number of successes in independent trials. *The Annals of Mathematical Statistics*, 36(4):1272–1278.

Scott, S. (2010). A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26:639 – 658.

Seznec, J., Locatelli, A., Carpentier, A., Lazaric, A., and Valko, M. (2019). Rotting bandits are no harder than stochastic ones. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2564–2572.

Takemori, S., Umeda, Y., and Gopalan, A. (2024). Model-based best arm identification for decreasing bandits. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *Proceedings of Machine Learning Research*, pages 1567–1575. PMLR.

Tang, W. and Tang, F. (2023). The Poisson Binomial Distribution—Old and New. *Statistical Science*, 38(1):108 – 119.

Tekin, C. and Liu, M. (2012). Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58(8):5588–5611.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Thornton, C., Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2013). Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *ACM international conference on Knowledge discovery and data mining (SIGKDD)*, pages 847–855.

Trovò, F., Paladino, S., Restelli, M., and Gatti, N. (2020). Sliding-window thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*, 68:311–364.

Wang, Y. H. (1993). On the number of successes in independent trials. *Statistica Sinica*, 3(2):295–312.

Waring, J., Lindvall, C., and Umeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial intelligence in medicine*, 104:101822.

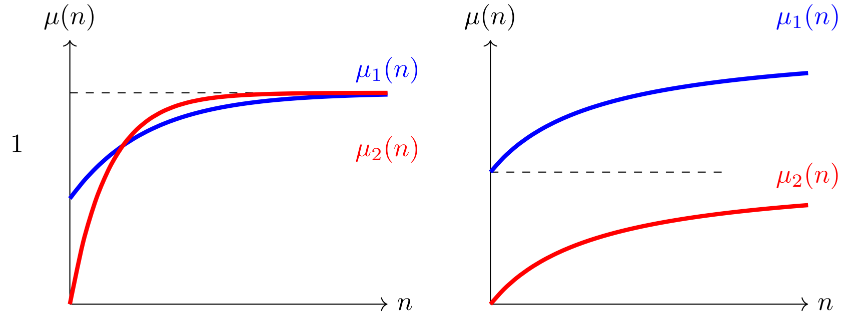
Xu, J., Wu, Y., Wang, Y., Wang, C., and Zheng, Z. (2023). Online experiments with diminishing marginal effects. *Available at SSRN 4640583*.

Checklist

1. For all models and algorithms presented, check if you include:
    (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes (Section 4).
    (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes (Section 4, 6, Appendix G).
    (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes (attached).
2. For any theoretical claim, check if you include:
    (a) Statements of the full set of assumptions of all theoretical results. Yes (Sections 5, 6).
    (b) Complete proofs of all theoretical results. Yes (Appendix C).
    (c) Clear explanations of any assumptions. Yes (Section 3).
3. For all figures and tables that present empirical results, check if you include:
    (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes (attached).
    (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes (Appendix F).
    (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes (Section 7).
    (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes (Appendix E)
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
    (a) Citations of the creator If your work uses existing assets. Yes (attached)
    (b) The license information of the assets, if applicable. Yes (attached)
    (c) New assets either in the supplemental material or as a URL, if applicable. Yes (attached)
    (d) Information about consent from data providers/curators. Not Applicable
    (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
    (a) The full text of instructions given to participants and screenshots. Not Applicable
    (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
    (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

SUPPLEMENTARY MATERIALS

A COMPARISON WITH THE `R-ED-UCB` ALGORITHM



(a) (b)

Figure 6: Different environments for the SRRB problem.

In this appendix, provide and analyze two SRRB instances to highlight the advantages and disadvantages of the proposed algorithms when compared with the optimistic algorithm `R-ed-UCB` Metelli et al. (2022) designed for SRRB settings. We recall the regret result provided by Metelli et al. (2022) for `R-ed-UCB`:

Theorem A.1 (Theorem 4.4, Metelli et al. (2022))**.** *`R-ed-UCB` with a suitable exploration index (see Metelli et al. (2022)) $\alpha > 2$, and $\epsilon \in (0, 1/2)$ suffers an expected regret for every $q \in [0, 1]$ bounded as:*

$$R(\texttt{R-ed-UCB}, T) \leqslant O\left(\frac{K}{\epsilon} T^{\frac{2}{3}} (\alpha \log T)^{\frac{1}{3}} + \frac{KT^q}{1-2\epsilon} \Upsilon\left(\left\lceil (1-2\epsilon)\frac{T}{K}\right\rceil, q\right)\right), \tag{7}$$

where $\Upsilon(M, q) := \sum_{l=1}^{M-1} \max_{i \in [\![K]\!]} \{\gamma_i(l)^q\}$ is a complexity index depending on the expected rewards.

First Instance. We start with an instance in which `R-ed-UCB` succeeds in delivering a sublinear regret, while our algorithms may fail. We define the expected reward functions as follows: $\mu_1(n) = 1 - 2^{-n}$ and $\mu_2(n) = 1 - 2^{-2n+2}$ (Figure 6a). It is not possible to find a σ, so that $\overline{\Delta}(\sigma, T) \geqslant \overline{\Delta}$, with $\overline{\Delta} > 0$ and not vanishing with T. This is visible since $\max_{\sigma \leqslant T} \overline{\Delta}(\sigma, T) \leqslant \frac{5}{6T}$. Therefore, we cannot guarantee that our algorithms provide a sublinear regret. Conversely, by using the definition of Υ, for $q \in [0, 1]$, we have:

$$\Upsilon\left(\left\lceil (1-2\epsilon)\frac{T}{K}\right\rceil, q\right) = \sum_{n=1}^{\lceil (1-2\epsilon)\frac{T}{K}\rceil} \max_{y \in \{1,2\}} \{e^{-y\lambda n} - e^{-y\lambda(n+1)}\}^q \leqslant \frac{3}{4} + \frac{1}{q \log 2},$$

which implies that Theorem A.1 provides a regret of order $O(T^{2/3} + T^q/q)$ for the `R-ed-UCB` algorithm which is sublinear for every $q < 1$ and, selecting $q = 1/\log T$ we obtain the best rate $O(T^{2/3} + \log T)$.

Second Instance. The second instance is designed so that our algorithms provide sublinear regret, while `R-ed-UCB` fails. We define the expected reward functions as: $\mu_1(n) = 1 - \frac{2^{\lambda-1}}{(n+1)^\lambda}$, and $\mu_2(n) = \frac{1}{2} - \frac{2^{\lambda-1}}{(n+1)^\lambda}$, where $\lambda \in [0, 1]$ is an arbitrary parameter (Figure 6b). With $\sigma = 2$ and $\overline{\Delta}(2, T) = \overline{\mu}_1(2) - \frac{1}{2}$, that is independent from T, since $\overline{\mu}_1(2) > \frac{1}{2} \geqslant \overline{\mu}_2(T)$ for all possible time horizons T. The Υ factor of this setting for every $q \in [0, 1]$ is given by:

$$\Upsilon\left(\left\lceil (1-2\epsilon)\frac{T}{K}\right\rceil, q\right) = \sum_{n=1}^{\lceil (1-2\epsilon)\frac{T}{K}\rceil} \left(\frac{2^{\lambda-1}}{(n+1)^\lambda} - \frac{2^{\lambda-1}}{(n+2)^\lambda}\right)^q \geqslant O\begin{cases} \lambda^q & \text{if } q(\lambda+1) > 1 \\ \lambda^{\frac{1}{\lambda+1}} \log T & \text{if } q(\lambda+1) = 1 \\ \lambda^q T^{1-q(\lambda+1)} & \text{otherwise} \end{cases}.$$

We prove in Appendix A.1, that the optimal choice of q is $1/(\lambda+1)$, according to Theorem A.1, this leads to a regret of order $O(T^{2/3} + (\lambda T)^{\frac{1}{\lambda+1}})$ for `R-ed-UCB`. Thus, for instance, choosing $\lambda = 1/3$, `R-ed-UCB` attains a regret bound of order $O(T^{3/4})$ while our approaches succeed to achieve an instance-dependent $O(\log(T))$ regret.

A.1 Detailed Computations for the Instances of Appendix A

First Instance. Let us upper bound the complexity index:

$$\begin{aligned}
\Upsilon\left(\left\lceil(1-2\epsilon)\frac{T}{K}\right\rceil, q\right) &= \sum_{n=1}^{\lceil(1-2\epsilon)\frac{T}{K}\rceil} \max\{2^{-n} - 2^{-(n+1)}, 2^{-2n+2} - 2^{-2(n+1)+2}\}^q \\
&\leqslant \frac{3}{4} + \sum_{n=2}^{\lceil(1-2\epsilon)\frac{T}{K}\rceil} (2^{-n} - 2^{-(n+1)} + 2^{-2n+2} - 2^{-2(n+1)+2})^q \\
&= \frac{3}{4} + \sum_{n=2}^{\lceil(1-2\epsilon)\frac{T}{K}\rceil} (2^{-1-2n}(6+2^n))^q \\
&\leqslant \frac{3}{4} + \sum_{n=2}^{\lceil(1-2\epsilon)\frac{T}{K}\rceil} \left(\frac{5}{4}2^{-n}\right)^q \\
&\leqslant \frac{3}{4} + \int_{n=2}^{+\infty} (2^{(-1-2n)}(6+2^n))^q \mathrm{d}n \\
&= \frac{3}{4} + \frac{1}{q\log 2}\left(\frac{5}{8}\right)^q \leqslant \frac{3}{4} + \frac{1}{q\log 2},
\end{aligned}$$

where we bounded the maximum with the sum, performed some algebraic bounds, and bounded the summation with the integral. For this instance, we compute the average expected rewards, for every $n \in [\![N]\!]$:

$$\overline{\mu}_1(n) = 1 - \frac{1}{n}\left(1 - 2^{-n}\right), \tag{8}$$

$$\overline{\mu}_1(n) = 1 - \frac{4}{3n}\left(1 - 2^{-2n}\right). \tag{9}$$

We will show that:

$$\inf_{T\in\mathbb{N}} \max_{\sigma\in[\![T]\!]} \overline{\Delta}(\sigma, T) = 0. \tag{10}$$

Let us fix $T \geqslant 1$, we have:

$$\begin{aligned}
\max_{\sigma\leqslant T} \overline{\Delta}(\sigma, T) &= \max_{\sigma\leqslant T}\left\{1 - \frac{1}{\sigma}\left(1-2^{-\sigma}\right) - 1 + \frac{4}{3T}\left(1-2^{-2T}\right)\right\} &(11)\\
&= 1 - \frac{1}{T}\left(1-2^{-T}\right) - 1 + \frac{4}{3T}\left(1-2^{-2T}\right) &(12)\\
&\leqslant \frac{1}{3T} + \frac{2^{-T}}{T} \leqslant \frac{5}{6T}, &(13)
\end{aligned}$$

being the function $-\frac{1}{\sigma}\left(1-2^{-\sigma}\right)$ non-decreasing in σ.

Second Instance. Let us lower bound the complexity index:

$$\begin{aligned}
\Upsilon\left(\left\lceil(1-2\epsilon)\frac{T}{K}\right\rceil, q\right) &= \sum_{n=1}^{\lceil(1-2\epsilon)\frac{T}{K}\rceil} \left(\frac{2^{\lambda-1}}{(n+1)^\lambda} - \frac{2^{\lambda-1}}{(n+2)^\lambda}\right)^q \\
&\geqslant \sum_{n=1}^{\lceil(1-2\epsilon)\frac{T}{K}\rceil} \left(\frac{2^{\lambda-1}\lambda}{(n+2)^{\lambda+1}}\right)^q,
\end{aligned}$$

where we used $\frac{2^{\lambda-1}}{(n+1)^\lambda} - \frac{2^{\lambda-1}}{(n+2)^\lambda} \geqslant \min_{x\in[n,n+1]} \frac{\partial}{\partial x}\left(1 - \frac{2^{\lambda-1}}{(x+1)^\lambda}\right) = \frac{2^{\lambda-1}\lambda}{(n+2)^{\lambda+1}}$. For $q(\lambda+1) > 1$, we proceed as follows:[15]

$$\sum_{n=1}^{\lceil (1-2\epsilon)\frac{T}{K} \rceil} \left(\frac{2^{\lambda-1}\lambda}{(n+2)^{\lambda+1}} \right)^q \geqslant 2^{q(\lambda-1)} 3^{-q(\lambda+1)} \lambda^q = O(\lambda^q).$$

Instead, for $q(\lambda+1) = 1$, we bound the summation with the integral:

$$\begin{aligned} \sum_{n=1}^{\lceil (1-2\epsilon)\frac{T}{K} \rceil} \left(\frac{2^{q(\lambda-1)}\lambda^q}{n+2} \right)^q &\geqslant \int_{n=1}^{(1-2\epsilon)\frac{T}{K}} \left(\frac{2^{q(\lambda-1)}\lambda^q}{n+2} \right) \mathrm{d}n \\ &\geqslant 2^{q(\lambda-1)}\lambda^q \log\left((1-2\epsilon)\frac{T}{K} - \frac{2}{3} \right) = O(\lambda^q \log T). \end{aligned}$$

Finally, for $q(\lambda+1) < 1$, we still bound the summation with the integral:

$$\begin{aligned} \sum_{n=1}^{\lceil (1-2\epsilon)\frac{T}{K} \rceil} \left(\frac{2^{\lambda-1}\lambda}{(n+2)^{\lambda+1}} \right)^q &\geqslant \int_{n=1}^{(1-2\epsilon)\frac{T}{K}} \left(\frac{2^{\lambda-1}\lambda}{(n+2)^{\lambda+1}} \right)^q \mathrm{d}n \\ &\geqslant \frac{2^{q(\lambda-1)}\lambda^q}{1-q(\lambda+1)} \left(\left((1-2\epsilon)\frac{T}{K} \right)^{1-q(\lambda+1)} - 3^{1-q(\lambda+1)} \right) \\ &= O(\lambda^q T^{1-q(\lambda+1)}). \end{aligned}$$

Now, recalling that the instance-dependent component of the regret of Theorem A.1 is in the order of $T^q \Upsilon\left(\left\lceil (1-2\epsilon)\frac{T}{K} \right\rceil, q \right)$, we have for the three cases the optimal choice of q that minimizes the regret:

$$\begin{aligned} q(\lambda+1) > 1 &\implies q \downarrow \frac{1}{\lambda+1} \implies O\left(\lambda^{\frac{1}{\lambda+1}} T^{\frac{1}{\lambda+1}} \right); \\ q(\lambda+1) = 1 &\implies q = \frac{1}{\lambda+1} \implies O\left(\lambda^{\frac{1}{\lambda+1}} T^{\frac{\lambda}{\lambda+1}} \log T \right); \\ q(\lambda+1) < 1 &\implies q \uparrow \frac{1}{\lambda+1} \implies O\left(\lambda^{\frac{1}{\lambda+1}} T^{\frac{1}{\lambda+1}} \right). \end{aligned}$$

Thus, we have that the bound of the instance-dependent component of the regret is at least $O\left(T^{\frac{1}{\lambda+1}} \right)$.

---

[15] We use big-O notation to highlight the dependences on $\lambda \to 0$ and $T \to +\infty$.

B PROOFS AND DERIVATIONS

In this appendix, we provide the complete proofs and derivations we have omitted in the main paper.

B.1 Proofs of Section 5

Lemma 5.1 (Expected Number of Pulls Bound for `ET-Beta-SWTS`). *Let $T \in \mathbb{N}$ be the learning horizon, $\tau \in [\![T]\!]$ the window size, $\Gamma \in [\![0, T]\!]$ the forced exploration parameter, for the* `ET-Beta-SWTS` *algorithm it holds for every $i \in [\![K]\!]$ and free parameter $\omega \in [\![0, T]\!]$ that:*

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mu}}[N'_{i,T}] \leqslant 1 + \Gamma + \frac{\omega T}{\tau} + \frac{T}{\tau} \\ + \mathbb{E}_{\boldsymbol{\mu}}\Bigg[\sum_{t \in \mathcal{T}_i} \mathbb{1}\Big\{p^i_{i,t,\tau} > \frac{1}{T},\ I_t = i,\ N_{i,t,\tau} \geqslant \omega\Big\}\Bigg] \\ + \mathbb{E}_{\boldsymbol{\mu}}\Bigg[\sum_{t \in \mathcal{T}_i} \Bigg(\frac{1}{p^i_{i^*(t),t,\tau}} - 1\Bigg) \mathbb{1}\left\{I_t = i^*(t)\right\}\Bigg].\end{aligned}$$

Proof. We will prove a more general result that will be useful later, in particular we will prove that for every $l \geqslant 0$, defining $\mathcal{T}_i := \{t \in [\![K\Gamma + l + 1, T]\!], i \neq i^*(t)\}$ the following statement holds:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mu}}[N'_{i,T}] \leqslant 1 + \Gamma + \frac{\omega T}{\tau} + \frac{T}{\tau} + l \\ + \mathbb{E}_{\boldsymbol{\mu}}\Bigg[\sum_{t \in \mathcal{T}_i} \mathbb{1}\Big\{p^i_{i,t,\tau} > \frac{1}{T},\ I_t = i,\ N_{i,t,\tau} \geqslant \omega\Big\}\Bigg] \\ + \mathbb{E}_{\boldsymbol{\mu}}\Bigg[\sum_{t \in \mathcal{T}_i} \Bigg(\frac{1}{p^i_{i^*(t),t,\tau}} - 1\Bigg) \mathbb{1}\left\{I_t = i^*(t)\right\}\Bigg].\end{aligned} \tag{14}$$

We define the event $E_i(t) := \{\theta_{i,t,\tau} \leqslant y_{i,t}\}$. Thus, the following holds:

$$\mathbb{E}[N'_{i,T}] = \sum_{t=1}^{T} \mathbb{P}(I_t = i, i \neq i^*(t)) \leqslant \Gamma + l + \underbrace{\frac{T}{\tau}}_{\text{(X)}} + \underbrace{\sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i^{\complement}(t))}_{\text{(A)}} + \underbrace{\sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i(t))}_{\text{(B)}}. \tag{15}$$

Term (X) arise due to the forced play whenever $N_{i,t,\tau} = 0$. Let us face term (A):

$$\begin{aligned}\text{(A)} &\leqslant \sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i^{\complement}(t), N_{i,t,\tau} \leqslant \omega) + \sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i^{\complement}(t), N_{i,t,\tau} \geqslant \omega) &(16)\\ &\leqslant \sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, N_{i,t,\tau} \leqslant \omega) + \sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i^{\complement}(t), N_{i,t,\tau} \geqslant \omega) &(17)\\ &= \mathbb{E}\left[\sum_{t \in \mathcal{T}_i} \mathbb{1}\left\{I_t = i, N_{i,t,\tau} \leqslant \omega\right\}\right] + \sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i^{\complement}(t), N_{i,t,\tau} \geqslant \omega) &(18)\\ &\leqslant \mathbb{E}\left[\underbrace{\sum_{t=1}^{T} \mathbb{1}\left\{I_t = i, N_{i,t,\tau} \leqslant \omega\right\}}_{\text{(C)}}\right] + \sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i^{\complement}(t), N_{i,t,\tau} \geqslant \omega). &(19)\end{aligned}$$

Observe that (C) can be bounded by Lemma C.13. Thus, the above inequality can be rewritten as:

$$\text{(A)} \leqslant \frac{\omega T}{\tau} + \underbrace{\sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i^{\complement}(t), N_{i,t,\tau} \geqslant \omega)}_{\text{(D)}}. \tag{20}$$

We now focus on the term (D). Defining $\mathcal{T}'_i := \{t \in \mathcal{T}_i : 1 - \mathbb{P}(\theta_{i,t,\tau} \leqslant y_{i,t} \mid \mathcal{F}_{t-1}) > \frac{1}{T}, N_{i,t,\tau} \geqslant \omega\}$ and

$\mathcal{T}_i'' := \{t \in \mathcal{T}_i : 1 - \mathbb{P}(\theta_{i,t,\tau} \leqslant y_{i,t} \mid \mathcal{F}_{t-1}) \leqslant \frac{1}{T}, N_{i,t,\tau} \geqslant \omega\}$ we obtain:

$$\sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i^\complement(t), N_{i,t,\tau} \geqslant \omega) = \mathbb{E}\left[\sum_{t \in \mathcal{T}_i} \mathbb{1}\{I_t = i, E_i^\complement(t), N_{i,t,\tau} \geqslant \omega\}\right] \tag{21}$$

$$= \mathbb{E}\left[\sum_{t \in \mathcal{T}_i'} \mathbb{1}\{I_t = i, E_i(t)^\complement\}\right] + \mathbb{E}\left[\sum_{t \in \mathcal{T}_i''} \mathbb{1}\{I_t = i, E_i(t)^\complement\}\right] \tag{22}$$

$$\leqslant \mathbb{E}\left[\sum_{t \in \mathcal{T}_i'} \mathbb{1}\{I_t = i\}\right] + \mathbb{E}\left[\sum_{t \in \mathcal{T}_i''} \mathbb{1}\{E_i(t)^\complement\}\right] \tag{23}$$

$$\leqslant \mathbb{E}\left[\sum_{t \in \mathcal{T}_i} \mathbb{1}\left\{1 - \mathbb{P}(\theta_{i,t,\tau} \leqslant y_{i,t} \mid \mathcal{F}_{t-1}) > \frac{1}{T}, N_{i,t,\tau} \geqslant \omega, I_t = i\right\}\right] + \sum_{t=1}^{T} \frac{1}{T}. \tag{24}$$

Now we focus on term (B). We have:

$$\sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i(t)) = \sum_{t \in \mathcal{T}_i} \mathbb{E}\left[\underbrace{\mathbb{P}(I_t = i, E_i(t) \mid \mathcal{F}_{t-1})}_{\text{(E)}}\right]. \tag{25}$$

In order to bound (B) we need to bound (E). Let $i_t' = \operatorname{argmax}_{i \neq i*(t)} \theta_{i,t,\tau}$. Then, we have:

$$\begin{aligned}\mathbb{P}\left(I_t = i^*(t), E_i(t) \mid \mathcal{F}_{t-1}\right) &\geqslant \mathbb{P}\left(i_t' = i, E_i(t), \theta_{i*(t),t,\tau} > y_{i,t} \mid \mathcal{F}_{t-1}\right) \\ &= \mathbb{P}\left(\theta_{i*(t),t,\tau} > y_{i,t} \mid \mathcal{F}_{t-1}\right) \mathbb{P}\left(i_t' = i, E_i(t) \mid \mathcal{F}_{t-1}\right) \\ &\geqslant \frac{p^i_{i*(t),t,\tau}}{1 - p^i_{i*(t),t,\tau}} \mathbb{P}\left(I_t = i, E_i(t) \mid \mathcal{F}_{t-1}\right),\end{aligned}$$

where in the first equality we used the fact that $\theta_{i*(t),t,\tau}$ is conditionally independent of i_t' and $E_i(t)$ given $\mathcal{F}_{t-1}$. In the second inequality, we used the fact that:

$$\mathbb{P}\left(I_t = i, E_i(t) \mid \mathcal{F}_{t-1}\right) \leqslant \left(1 - \mathbb{P}\left(\theta_{i*(t),t,\tau} > y_{i,t} \mid \mathcal{F}_{t-1}\right)\right) \mathbb{P}\left(i_t' = i, E_i(t) \mid \mathcal{F}_{t-1}\right)$$

which is true since $\{I_t = i\} \cap E_i(t) \subseteq \{i_t' = i\} \cap E_i(t) \cap \left\{\theta_{i*(t),t,\tau} \leqslant y_{i,t}\right\}$, and the events $\{i_t' = i\} \cap E_i(t)$ and $\left\{\theta_{i*(t),t,\tau} \leqslant y_{i,t}\right\}$ are conditionally independent given $\mathcal{F}_{t-1}$. Therefore, we have:

$$\begin{aligned}\mathbb{P}\left(I_t = i, E_i(t) \mid \mathcal{F}_{t-1}\right) &\leqslant \left(\frac{1}{p^i_{i*(t),t,\tau}} - 1\right) \mathbb{P}\left(I_t = i^*(t), E_i(t) \mid \mathcal{F}_{t-1}\right) \\ &\leqslant \left(\frac{1}{p^i_{i*(t),t,\tau}} - 1\right) \mathbb{P}\left(I_t = i^*(t) \mid \mathcal{F}_{t-1}\right),\end{aligned}$$

substituting, we obtain:

$$\sum_{t \in \mathcal{T}_i} \mathbb{E}[\mathbb{P}(I_t = i, E_i(t) \mid \mathcal{F}_{t-1})] \leqslant \mathbb{E}\left[\sum_{t \in \mathcal{T}_i} \left(\frac{1}{p^i_{i*(t),t,\tau}} - 1\right) \mathbb{P}\left(I_t = i^*(t) \mid \mathcal{F}_{t-1}\right)\right] \tag{26}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\sum_{t \in \mathcal{T}_i} \left(\frac{1}{p^i_{i*(t),t,\tau}} - 1\right) \mathbb{1}\left\{I_t = i^*(t)\right\} \mid \mathcal{F}_{t-1}\right]\right] \tag{27}$$

$$= \mathbb{E}\left[\sum_{t \in \mathcal{T}_i} \left(\frac{1}{p^i_{i*(t),t,\tau}} - 1\right) \mathbb{1}\left\{I_t = i^*(t)\right\}\right]. \tag{28}$$

The statement follows by summing all the terms. ☐

Lemma 5.2 (PB-Bin Stochastic Dominance). *Let $j \in \mathbb{N}$, $\mathrm{PB}(\underline{\mu}_{i*(t)}(j))$ be a Poisson-Binomial distribution with parameters $\underline{\mu}_{i*(t)}(j) = (\mu_{i*(t),1}, \ldots, \mu_{i*(t),j})$, and $\mathrm{Bin}(j, x)$ be a binomial distribution of j trials and success probability $0 \leqslant x \leqslant \frac{1}{j}\sum_{l=1}^{j} \mu_{i*(t),l} = \overline{\mu}_{i*(t),j}$. Then, it holds that:*

$$\mathop{\mathbb{E}}_{S_{i*(t),t} \sim \mathrm{PB}(\underline{\mu}_{i*(t)}(j))}\left[\frac{1}{p^i_{i*(t),t}} \middle| N_{i*(t),t} = j\right]$$

$$\leqslant \mathop{\mathbb{E}}_{S_{i*(t),t} \sim \mathrm{Bin}(j,x)} \left[\frac{1}{p^i_{i*(t),t}} \middle| N_{i*(t),t} = j \right].$$

Proof. Let us assume, without loss of generality, the arm 1 is the optimal arm at round $t \in [\![T]\!]$. Let $N_{1,t} = j$, $S_{1,t} = s$. Then, as shown by Agrawal and Goyal (2012), $p^i_{1,t}$ can be written as:

$$p^i_{1,t} = \mathbb{P}(\theta_{1,t} > y_{i,t}) = F^B_{j+1,y_{i,t}}(s),$$

where $F^B_{j+1,y_{i,t}}(s)$ is the cumulative distribution function of a Binomial random variable after $j+1$ Bernoulli trials each with probability of success $y_{i,t}$ evaluated in s. Our goal is to prove the inequality:

$$\mathop{\mathbb{E}}_{X' \sim \mathrm{PB}(\underline{\mu}_1(j))} \left[\frac{1}{F^B_{j+1,y_{i,t}}(X')} \right] \leqslant \mathop{\mathbb{E}}_{X \sim \mathrm{Bin}(j,x)} \left[\frac{1}{F^B_{j+1,y_{i,t}}(X)} \right]. \tag{29}$$

We notice that the probability mass function of a binomial distribution is discrete log-concave (Lemma C.11). Thus, let $Y \sim \mathrm{Bin}(j+1, y_{i,t})$, we have that for $i \in [\![0, j-1]\!]$ it holds that:

$$p_Y(i+1)^2 \geqslant p_Y(i) p_Y(i+2). \tag{30}$$

By Lemma C.12 used with $\alpha = 1$ and $r = +\infty$, and q being the probability mass function of the binomial distribution (precisely, $q(x) = p(-x)$ in Lemma C.12), we find that the CDF of the binomial is discrete log-concave on $\mathbb{Z}$ too. Indeed, according to Lemma C.12, if the probability mass function of an integer-valued random variable is discrete log-concave as a function on $\mathbb{Z}$, then the corresponding CDF ($F^B_{j+1,y_{i,t}}$ in our notation) is also discrete log-concave as a function on $\mathbb{Z}$. Thus, omitting superscripts and subscripts, $1/F$ is discrete log-convex on the set $x \in S := [\![0, j+1]\!]$, i.e., for $x \in [\![0, j-1]\!]$:

$$\left(\frac{1}{F(x+1)} \right)^2 \leqslant \frac{1}{F(x+2)} \frac{1}{F(x)}, \tag{31}$$

or, equivalently:

$$\frac{1}{F(x+1)} \leqslant \left(\frac{1}{F(x+2)} \right)^{\frac{1}{2}} \left(\frac{1}{F(x)} \right)^{\frac{1}{2}}. \tag{32}$$

Using the AM-GM inequality, we obtain:

$$\frac{1}{F(x+1)} < \frac{1}{2} \left(\frac{1}{F(x+2)} \right) + \frac{1}{2} \left(\frac{1}{F(x)} \right). \tag{33}$$

Notice the inequality is strict since the AM-GM inequality holds with equality only when all elements are equal and this is not our case since $\frac{1}{F(x+2)} < \frac{1}{F(x)}$ for $x \in [\![0, j-1]\!]$, and hence $\frac{1}{F(x+2)} \neq \frac{1}{F(x)}$. Thus, we have proved that $1/F$ is strictly discrete convex on S. Therefore, using Lemma C.10 we obtain that for every j, being the number of trials for both the Poisson-binomial and the binomial distributions, the expected value of the term of our interest for a Poisson-binomial distribution with a certain mean of the probabilities of the success at each trial, namely $\overline{\mu}_{1,j} = \frac{1}{j} \sum_{l=1}^{j} \mu_{1,l}$, is always smaller than the one of a binomial distribution where each Bernoulli trial has a probability of success equal to $\overline{\mu}_{1,j}$. More formally:

$$\mathop{\mathbb{E}}_{X' \sim \mathrm{PB}(\underline{\mu}_1(j))} \left[\frac{1}{F^B_{j+1,y_{i,t}}(X')} \right] \leqslant \mathop{\mathbb{E}}_{X'' \sim \mathrm{Bin}(j,\overline{\mu}_{1,j})} \left[\frac{1}{F^B_{j+1,y_{i,t}}(X'')} \right]. \tag{34}$$

To show Equation (29) for any j such that $\overline{\mu}_{1,j} \geqslant x$, we need to prove that the expected value of $1/F^B_{j+1,y_{i,t}}$ considered for a binomial process with mean $\overline{\mu}_{1,j}$ is smaller than the expected value of $1/F^B_{j+1,y_{i,t}}$ for a binomial distribution with mean $x \leqslant \overline{\mu}_{1,j}$. We apply Lemma C.3 stating that for a non-negative random variable (like ours $1/F^B_{j+1,y_{i,t}}$), the expected value can be computed as:

$$\mathbb{E}_{\mathrm{Bin}(j,\overline{\mu}_{1,j})} \left[\frac{1}{F^B_{j+1,y_{i,t}}} \right] = \int_0^{+\infty} \mathbb{P} \left(\frac{1}{F^B_{j+1,y_{i,t}}} > y \right) \mathrm{d}y. \tag{35}$$

Let $X'' \sim \mathrm{Bin}(j, \overline{\mu}_{1,j})$. Thus, we have:

$$\mathbb{P} \left(\frac{1}{F^B_{j+1,y_{i,t}}} > y \right) = \mathbb{P}(X'' = 0) + \mathbb{P}(X'' = 1) + \ldots + \mathbb{P} \left(X'' = \left(\frac{1}{F^B_{j+1,y_{i,t}}} \right)^{-1} (y) - 1 \right) \tag{36}$$

$$= \mathbb{P}\left(X'' \leqslant \underbrace{\left(\frac{1}{F^{B}_{j+1,y_{i,t}}}\right)^{-1}(y) - 1}_{=:\, k_j(y)} \right), \tag{37}$$

and the same goes for $X \sim \mathrm{Bin}(j, x)$:

$$\mathbb{P}\left(\frac{1}{F^{B}_{j+1,y_i}} > y\right) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \ldots + \mathbb{P}\left(X = \left(\frac{1}{F^{B}_{j+1,y_{i,t}}}\right)^{-1}(y) - 1\right) \tag{38}$$

$$= \mathbb{P}\left(X \leqslant \underbrace{\left(\frac{1}{F^{B}_{j+1,y_{i,t}}}\right)^{-1}(y) - 1}_{=:\, k_j(y)} \right), \tag{39}$$

where the inverse is formally defined as follows:

$$\left(\frac{1}{F^{B}_{j+1,y_{i,t}}}\right)^{-1}(y) := \min\left\{ s \in [\![0, j]\!] \,:\, y \geqslant \frac{1}{F^{B}_{j+1,y_{i,t}}(s)} \right\}. \tag{40}$$

Notice, that whenever such s in the above definition does not exist, we will have that $\mathbb{P}\left(X < \left(\frac{1}{F^{B}_{j+1,y_{i,t}}}\right)^{-1}(y)\right) = \mathbb{P}\left(X'' < \left(\frac{1}{F^{B}_{j+1,y_{i,t}}}\right)^{-1}(y)\right) = 0$, so that those values does not contribute to the integrals. Thus, we need to prove:

$$\mathbb{E}_{\mathrm{Bin}(j,\overline{\mu}_{1,j})}\left[\frac{1}{F^{B}_{j+1,y_{i,t}}}\right] = \int_0^{+\infty} \mathbb{P}(X'' \leqslant k_j(y))\mathrm{d}y \leqslant \int_0^{+\infty} \mathbb{P}(X \leqslant k_j(y))\mathrm{d}y = \mathbb{E}_{\mathrm{Bin}(j,x)}\left[\frac{1}{F^{B}_{j+1,y_{i,t}}}\right]. \tag{41}$$

A sufficient condition to ensure that the condition in Equation (41) holds is that:

$$\mathbb{P}(X'' > m) \geqslant \mathbb{P}(X > m), \quad \forall m \in \mathbb{R}. \tag{42}$$

Indeed, we have

$$\int_0^{+\infty} \mathbb{P}(X'' \leqslant k_j(y))\mathrm{d}y \leqslant \int_0^{+\infty} \mathbb{P}(X \leqslant k_j(y))\mathrm{d}y \tag{43}$$

$$\iff \int_0^{+\infty} (1 - \mathbb{P}(X'' > k_j(y)))\mathrm{d}y \leqslant \int_0^{+\infty} (1 - \mathbb{P}(X > k_j(y)))\mathrm{d}y \tag{44}$$

$$\iff \int_0^{+\infty} \mathbb{P}(X'' > k_j(y))\mathrm{d}y \geqslant \int_0^{+\infty} \mathbb{P}(X > k_j(y))\mathrm{d}y \tag{45}$$

$$\iff \int_0^{+\infty} \big(\mathbb{P}(X'' > k_j(y)) - \mathbb{P}(X > k_j(y))\big)\mathrm{d}y \geqslant 0. \tag{46}$$

Let us recall the concept of stochastic order (Boland et al. (2002, 2004); Marshall et al. (2011)) that is often useful in comparing random variables. For two random variables U and V, we say that U is *greater* than V in the usual stochastic order (or U *stochastically dominates* V), and we denote it with $U \geqslant_{\mathrm{st}} V$, when $\mathbb{P}(U > m) \geqslant \mathbb{P}(V > m)$ for every $m \in \mathbb{R}$. Thus, if we have that $X'' \geqslant_{\mathrm{st}} X$ we would have that also Equation (42) holds too. It has been shown by Boland et al. (2002) (Lemma C.6) that the condition for that to happen when X'' and X are binomial distributions with means μ'' and μ is that $\mu'' \geqslant \mu$. Thus, we have showed that for any j such that $\overline{\mu}_{1,j} \geqslant x$:

$$\mathbb{E}_{X' \sim \mathrm{PB}(\underline{\mu}_1(j))}\left[\frac{1}{F_{j+1,y_{i,t}}(X')}\right] \leqslant \mathbb{E}_{X'' \sim \mathrm{Bin}(j,\overline{\mu}_{1,j})}\left[\frac{1}{F_{j+1,y_{i,t}}(X'')}\right] \leqslant \mathbb{E}_{X \sim \mathrm{Bin}(j,\, x)}\left[\frac{1}{F_{j+1,y_{i,t}}(X)}\right]. \tag{47}$$

This concludes the proof. □

B.2 Additional Lemmas

Lemma B.1 (Expected Number of Pulls Bound for γ-`ET-SWGTS`)**.** *Let $T \in \mathbb{N}$ be the learning horizon, $\tau \in [\![T]\!]$ be the window size, $\Gamma \in [\![T]\!]$ be the forced exploration parameter, for the γ-`ET-SWGTS` algorithm the following holds* *for every free parameters* $\omega \in [\![0, T]\!]$ *and* $\epsilon_i > 0$*:*

$$\mathbb{E}[N'_{i,T}] \leqslant \Gamma + \frac{1}{\epsilon_i} + \frac{T}{\tau} + \frac{\omega T}{\tau} + \mathbb{E}\left[\sum_{t \in \mathcal{T}_i} \mathbb{1}\left\{p^i_{i,t,\tau} > \frac{1}{T\epsilon_i},\ I_t = i,\ N_{i,t,\tau} \geqslant \omega, \right\}\right]$$
$$+ \mathbb{E}\left[\sum_{t \in \mathcal{T}_i} \left(\frac{1}{p^i_{i^*(t),t,\tau}} - 1\right) \mathbb{1}\left\{I_t = i^*(t)\right\}\right].$$

Proof. We will prove a more general result that will be useful later, in particular we will prove that for every $l \geqslant 0$, defining $\mathcal{T}_i := \{t \in [\![K\Gamma + l + 1, T]\!], i \neq i^*(t)\}$ the following statement holds:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mu}}[N_{i,T}] &\leqslant \Gamma + \frac{1}{\epsilon_i} + \frac{\omega T}{\tau} + \frac{T}{\tau} + l \\ &+ \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t \in \mathcal{T}_i} \mathbb{1}\left\{p^i_{i,t,\tau} > \frac{1}{T},\ I_t = i,\ N_{i,t,\tau} \geqslant \omega\right\}\right] \\ &+ \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t \in \mathcal{T}_i} \left(\frac{1}{p^i_{i^*(t),t,\tau}} - 1\right) \mathbb{1}\left\{I_t = i^*(t)\right\}\right]. \end{aligned} \tag{48}$$

We define the event $E_i(t) := \{\theta_{i,t,\tau} \leqslant y_{i,t}\}$. Thus, the following holds:

$$\mathbb{E}_\tau[N'_{i,T}] = \sum_{t=1}^{T} \mathbb{P}(I_t = i, i \neq i^*(t)) \leqslant \Gamma + l + \underbrace{\frac{T}{\tau}}_{\text{(X)}} + \underbrace{\sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i^{\complement}(t))}_{\text{(A)}} + \underbrace{\sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i(t))}_{\text{(B)}}, \tag{49}$$

where (X) is the term arising given by the forced play whenever $N_{i,t,\tau} = 0$. Let us first face term (A):

$$\begin{aligned}\text{(A)} &\leqslant \sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i^{\complement}(t), N_{i,t,\tau} \leqslant \omega) + \sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i^{\complement}(t), N_{i,t,\tau} \geqslant \omega) && (50)\\ &\leqslant \sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, N_{i,t,\tau} \leqslant \omega) + \sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i^{\complement}(t), N_{i,t,\tau} \geqslant \omega) && (51)\\ &\leqslant \mathbb{E}\left[\sum_{t \in \mathcal{T}_i} \mathbb{1}\left\{I_t = i, N_{i,t,\tau} \leqslant \omega\right\}\right] + \sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i^{\complement}(t), N_{i,t,\tau} \geqslant \omega) && (52)\\ &\leqslant \mathbb{E}\left[\underbrace{\sum_{t=1}^{T} \mathbb{1}\left\{I_t = i, N_{i,t,\tau} \leqslant \omega\right\}}_{\text{(C)}}\right] + \sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i^{\complement}(t), N_{i,t,\tau} \geqslant \omega). && (53)\end{aligned}$$

Observe that (C) can be bounded by Lemma C.13. Thus, the above inequality can be rewritten as:

$$\text{(A)} \leqslant \frac{\omega T}{\tau} + \underbrace{\sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i^{\complement}(t), N_{i,t,\tau} \geqslant \omega)}_{\text{(D)}}. \tag{54}$$

We now focus on the term (D). Defining $\mathcal{T}'_i := \{t \in \mathcal{T}_i : 1 - \mathbb{P}(\theta_{i,t,\tau} \leqslant y_{i,t} \mid \mathcal{F}_{t-1}) > \frac{1}{T\epsilon_i}, N_{i,t,\tau} \geqslant \omega\}$ and $\mathcal{T}''_i := \{t \in \mathcal{T}_i : 1 - \mathbb{P}(\theta_{i,t,\tau} \leqslant y_{i,t} \mid \mathcal{F}_{t-1}) \leqslant \frac{1}{T\epsilon_i}, N_{i,t,\tau} \geqslant \omega\}$ we obtain:

$$\begin{aligned}\sum_{t \in \mathcal{T}_i} \mathbb{P}(I_t = i, E_i^{\complement}(t), N_{i,t,\tau} \geqslant \omega) &= \mathbb{E}\left[\sum_{t \in \mathcal{T}_i} \mathbb{1}\{I_t = i, E_i^{\complement}(t), N_{i,t,\tau} \geqslant \omega\}\right] && (55)\\ &= \mathbb{E}\left[\sum_{t \in \mathcal{T}'_i} \mathbb{1}\{I_t = i, E_i(t)^{\complement}\}\right] + \mathbb{E}\left[\sum_{t \in \mathcal{T}''_i} \mathbb{1}\{I_t = i, E_i(t)^{\complement}\}\right] && (56)\\ &\leqslant \mathbb{E}\left[\sum_{t \in \mathcal{T}'_i} \mathbb{1}\{I_t = i\}\right] + \mathbb{E}\left[\sum_{t \in \mathcal{T}''_i} \mathbb{1}\{E_i(t)^{\complement}\}\right] && (57)\end{aligned}$$

$$\leqslant \mathbb{E}\left[\sum_{t\in\mathcal{T}_i} \mathbb{1}\left\{1-\mathbb{P}(\theta_{i,t,\tau}\leqslant y_{i,t}\mid \mathcal{F}_{t-1}) > \frac{1}{T\epsilon_i}, N_{i,t,\tau}\geqslant \omega, I_t = i\right\}\right] + \sum_{t=1}^{T}\frac{1}{T\epsilon_i}. \tag{58}$$

Term (B) is bounded exactly as in the proof of Lemma 5.1. The statement follows by summing all the terms. ☐

B.3 Proofs of Section 6

Lemma 6.1 (Wald's Inequality for Rising Bandits)**.** *Let $\mathfrak{A}$ be an algorithm and $T \in \mathbb{N}$ be a learning horizon, it holds that: $R_{\boldsymbol{\mu}}(\mathfrak{A},T) \leqslant \sum_{i\neq i^*(T)} \Delta_i(T,1)\,\mathbb{E}_{\boldsymbol{\mu}}[N_{i,T}]$.*

Proof. We start with the usual definition of regret and proceed as follows:

$$R(\mathfrak{A},T) = T\overline{\mu}_{i^*(T)}(T) - \mathbb{E}\left[\sum_{t=1}^{T}\mu_{I_t}(N_{I_t,t})\right] \tag{59}$$

$$= \mathbb{E}\left[\sum_{t=1}^{T}\left(\mu_{i^*(T)}(t) - \mu_{I_t}(N_{I_t,t})\right)\right] \tag{60}$$

$$= \mathbb{E}\left[\sum_{t=1}^{T}\mu_{i^*(T)}(t) - \sum_{j=1}^{N_{i^*(T),T}}\mu_{i^*(T)}(j) - \sum_{i\neq i^*(T)}\sum_{j=1}^{N_{i,T}}\mu_i(j)\right] \tag{61}$$

$$= \mathbb{E}\left[\sum_{j=N_{i^*(T),T}+1}^{T}\mu_{i^*(T)}(j) - \sum_{i\neq i^*(T)}\sum_{j=1}^{N_{i,T}}\mu_i(j)\right] \tag{62}$$

$$\leqslant \mathbb{E}\left[\sum_{i\neq i^*(T)}\sum_{j=1}^{N_{i,T}}\left(\mu_{i^*(T)}(T) - \mu_i(j)\right)\right] \tag{63}$$

$$\leqslant \sum_{i\neq i^*(T)}\left(\mu_{i^*(T)}(T) - \mu_i(1)\right)\mathbb{E}\left[\sum_{j=1}^{N_{i,T}}1\right]. \tag{64}$$

The result follows from the definition of $\Delta_i(T,1)$. ☐

Theorem 6.2 (`ET-Beta-TS` Bound)**.** *Let $\sigma \in [\![\sigma_{\boldsymbol{\mu}}(T),T]\!]$. Under Assumptions 3.1 and 3.2, for the* `ET-Beta-TS` *algorithm, for every arm $i \in [\![K]\!]\backslash\{i^*(T)\}$, it holds for every $\epsilon \in (0,1]$, that:*[16]

$$\mathbb{E}_{\boldsymbol{\mu}}[N_{i,T}] \leqslant O\Bigg(\underbrace{\Gamma}_{(i)} + \underbrace{\frac{(1+\epsilon)\log(T)}{d(\overline{\mu}_i(T),\overline{\mu}_{i^*(T)}(\sigma))} + \frac{1}{\epsilon^2}}_{(ii)}$$
$$+ \underbrace{\sum_{j=\Gamma}^{\sigma}\frac{\delta_{\mathrm{TV}}(\mathrm{Bin}(j,\overline{\mu}_{i^*(T)}(j)),\mathrm{Bin}(j,\overline{\mu}_{i^*(T)}(\sigma))}{(1-\overline{\mu}_{i^*(T)}(\sigma))^{j+1}}}_{(iii)}\Bigg),$$

where $d(x,y) := x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$ is the KL divergence between Bernoulli distributions of parameters $x,y \in [0,1]$, $\delta_{\mathrm{TV}}(P,Q) := \sup_{A\in\mathcal{F}}|P(A)-Q(A)|$ is the total variation between P and Q, $\mathrm{Bin}(j,x)$ is the binomial distribution with j trials and parameter x.

Proof. In order to bound the expected value of the pulls of the suboptimal arm $i \neq i^*(T)$ at the time horizon T, we use Lemma 5.1, considering the forced exploration and imposing the window length $\tau = T$, $l = 0$. Formally, for `Beta-ET-TS`, we have:

$$\mathbb{E}[N_{i,T}] \leqslant \Gamma + 2 + \frac{\omega T}{T} + \underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+1}^{T}\mathbb{1}\left\{p^i_{i,t,T} > \frac{1}{T},\ I_t = i,\ N_{i,t}\geqslant\omega\right\}\right]}_{\text{(A)}}$$

---

[16] For the sake of presentation, with the Big-O notation, we are neglecting terms depending on $\mu_i(T)$ and $\mu_{i^*(T)}(\sigma)$, but not explicitly on T. The full expression is reported in Equation (112) in the appendix.

$$+\underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+1}^{T}\left(\frac{1}{p^{i}_{i^*(t),t,T}}-1\right)\mathbb{1}\left\{I_t=i^*(t)\right\}\right]}_{\text{(B)}}. \tag{65}$$

In the following, we consider, without loss of generality, the first arm to be the best, i.e., for every round $t \in [\![T]\!]$ we have $i^*(t) = i^*(T) = 1$ and introduce the following quantities, defined for some $\epsilon \in (0,1]$: $x_{i,t} = x_i \in (\overline{\mu}_i(T), \overline{\mu}_1(\sigma))$ such that $d(x_i, \overline{\mu}_1(\sigma)) = d(\overline{\mu}_i(T), \overline{\mu}_1(\sigma))/(1+\epsilon)$, $y_{i,t} = y_i \in (x_i, \overline{\mu}_1(\sigma))$ such that $d(x_i, y_i) = d(x_i, \overline{\mu}_1(\sigma))/(1+\epsilon) = d(\overline{\mu}_i(T), \overline{\mu}_1(\sigma))/(1+\epsilon)^2$, $\omega = \frac{\log(T)}{d(x_i,y_i)} = (1+\epsilon)^2 \frac{\log(T)}{d(\overline{\mu}_i(T),\overline{\mu}_1(\sigma))}$. We define $\hat{\mu}_{i,t,T} = \frac{S_{i,t}}{N_{i,t}}$.

We further decompose term $\mathcal{A}$ in two contributions:

$$\text{(A)} = \underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+1}^{T}\mathbb{1}\left\{p^{i}_{i,t,T} > \frac{1}{T}, \hat{\mu}_{i,t,T} \leqslant x_i, I_t = i,\ N_{i,t} \geqslant \omega\right\}\right]}_{\text{(A1)}}$$

$$+\underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+1}^{T}\mathbb{1}\left\{p^{i}_{i,t,T} > \frac{1}{T}, \hat{\mu}_{i,t,T} > x_i, I_t = i,\ N_{i,t} \geqslant \omega\right\}\right]}_{\text{(A2)}} \tag{66}$$

Term (A2) Focusing on (A2), let τ^i_j denote the round in which arm i is played for the j-th time:

$$\text{(A2)} \leqslant \mathbb{E}\left[\sum_{t=K\Gamma+1}^{T}\mathbb{1}\left\{\hat{\mu}_{i,t,T} > x_i, I_t = i, N_{i,t} \geqslant \omega\right\}\right] \tag{67}$$

$$\leqslant \mathbb{E}\left[\sum_{j=K\Gamma}^{T}\mathbb{1}\left\{\hat{\mu}_{i,\tau^i_j+1,T} > x_i\right\}\sum_{t=\tau^i_j+1}^{\tau^i_{j+1}}\mathbb{1}\left\{I_t = i\right\}\right] \tag{68}$$

$$\leqslant \sum_{j=\Gamma}^{T}\mathbb{P}(\hat{\mu}_{i,\tau^i_j+1,T} > x_i) \tag{69}$$

$$= \sum_{j=\Gamma}^{T}\mathbb{P}(\hat{\mu}_{i,\tau^i_j+1,T} - \mathbb{E}[\hat{\mu}_{i,\tau^i_j+1,T}] > x_i - \mathbb{E}[\hat{\mu}_{i,\tau^i_j+1,T}]) \tag{70}$$

$$\leqslant \sum_{j=\Gamma}^{T}\exp\left(-j d(x_i, \mathbb{E}[\hat{\mu}_{i,\tau^i_j+1,T}])\right) \tag{71}$$

$$\leqslant \sum_{j=0}^{T}\exp\left(-j d(x_i, \overline{\mu}_i(T))\right), \tag{72}$$

where in the last but one inequality, we have exploited the Chernoff-Hoeffding bound (Lemma C.1 with $\lambda = x_i - \mathbb{E}[\hat{\mu}_{i,\tau^i_j+1,T}]$), and in the last one, we observed that $\mathbb{E}[\hat{\mu}_{i,\tau^i_j+1,T}] \leqslant \overline{\mu}_i(T)$. Thus, we have:

$$\sum_{j=0}^{T}\exp\left(-j d(x_i, \overline{\mu}_i(T))\right) \leqslant 1 + \int_{j=0}^{+\infty}\exp\left(-j d(x_i, \overline{\mu}_i(T))\right)\mathrm{d}j \tag{73}$$

$$= 1 + \frac{1}{d(x_i, \overline{\mu}_i(T))} \tag{74}$$

$$\leqslant 1 + \frac{1}{2(x_i - \overline{\mu}_i(T))^2}, \tag{75}$$

having used Pinsker's inequality. Following Agrawal and Goyal (2017), we have the inequality:

$$x_i - \overline{\mu}_i(T) \geqslant \frac{\epsilon}{1+\epsilon} \cdot \frac{d(\overline{\mu}_i(T), \overline{\mu}_1(\sigma))}{\log\frac{\overline{\mu}_1(\sigma)(1-\overline{\mu}_i(T))}{\overline{\mu}_i(T)(1-\overline{\mu}_1(\sigma))}}. \tag{76}$$

From this, we have:

$$\sum_{s=0}^{T} \exp\left(-sd(x_i, \overline{\mu}_i(T))\right) \leqslant 1 + \left(\log \frac{\overline{\mu}_1(\sigma)(1-\overline{\mu}_i(T))}{\overline{\mu}_i(T)(1-\overline{\mu}_1(\sigma))}\right)^2 \cdot \frac{(1+\epsilon)^2}{2\epsilon^2} \cdot \frac{1}{d(\overline{\mu}_i(T), \overline{\mu}_1(\sigma))^2}. \tag{77}$$

Term (A1) Let us focus on the term (A1) of the regret.

$$\text{(A1)} \leqslant \mathbb{E}\left[\sum_{t=K\Gamma+1}^{T} \mathbb{1}\left\{\underbrace{p^i_{i,t,T} > \frac{1}{T}, N_{i,t} \geqslant \omega, \hat{\mu}_{i,t,T} \leqslant x_i}_{\text{(C)}}\right\}\right]. \tag{78}$$

We wish to evaluate if the condition (C) ever occurs. To this end, let $(\mathcal{F}_{t-1})_{t\in[\![T]\!]}$ be the canonical filtration. We have:

$$\begin{aligned}
&\mathbb{P}(\theta_{i,t,T} > y_i | \hat{\mu}_{i,t,T} \leqslant x_i, N_{i,t} \geqslant \omega, \mathcal{F}_{t-1}) = \\
&= \mathbb{P}\left(\text{Beta}\left(\hat{\mu}_{i,t,T}N_{i,t}+1, (1-\hat{\mu}_{i,t,T})N_{i,t}+1\right) > y_i | \hat{\mu}_{i,t,T} \leqslant x_i, N_{i,t} \geqslant \omega\right) && (79)\\
&\leqslant \mathbb{P}\left(\text{Beta}\left(x_i N_{i,t}+1, (1-x_i)N_{i,t}+1\right) > y_i | N_{i,t} \geqslant \omega\right) && (80)\\
&\leqslant F^B_{N_{i,t}+1,y_i}\left(x_i N_{i,t} | N_{i,t} \geqslant \omega\right) \leqslant F^B_{N_{i,t},y_i}\left(x_i N_{i,t} | N_{i,t} \geqslant \omega\right) && (81)\\
&\leqslant \exp\left(-(N_{i,t})d(x_i, y_i) | N_{i,t} \geqslant \omega\right) && (82)\\
&\leqslant \exp\left(-\omega d(x_i, y_i)\right), && (83)
\end{aligned}$$

where (82) follows from the generalized Chernoff-Hoeffding bounds (Lemma C.1) and (80) from the Beta-Binomial identity (Fact C.5). Equation (79) was derived by exploiting the fact that on the event $\hat{\mu}_{i,t,\tau} \leqslant x_i$ a sample from Beta $(x_i N_{i,t}+1, (1-x_i)N_{i,t}+1)$ is likely to be as large as a sample from $\text{Beta}(\hat{\mu}_{i,t,T}N_{i,t}+1, (1-\hat{\mu}_{i,t,T})N_{i,t}+1)$, as a $\text{Beta}(\alpha, \beta)$ random variable is stochastically dominated by $\text{Beta}(\alpha', \beta')$ if $\alpha' \geqslant \alpha$ and $\beta' \leqslant \beta$ (Lemma C.17), and the second inequality in (81) derives from Lemma C.16. Therefore, since $\omega = \frac{\log(T)}{d(x_i, y_i)}$, we have:

$$\mathbb{P}(\theta_{i,t,T} > y_i | \hat{\mu}_{i,t,T} \leqslant x_i, N_{i,t} \geqslant \omega, \mathbb{F}_{t-1}) \leqslant \frac{1}{T}, \tag{84}$$

in contradiction with condition (C), implying that (A1)= 0.

Term (B) For this term, We have:

$$\text{(B)} = \mathbb{E}\left[\sum_{t=K\Gamma+1}^{T} \left(\frac{1}{p^i_{1,t,T}} - 1\right) \mathbb{1}\left\{I_t = 1\right\}\right] \tag{85}$$

Let τ_j denote the time step at which arm 1 is played for the j-th time:

$$\begin{aligned}
\text{(B)} &\leqslant \sum_{j=\Gamma}^{T-1} \mathbb{E}\left[\frac{1-p^i_{1,\tau_j+1,T}}{p^i_{1,\tau_j+1,T}} \sum_{t=\tau_j+1}^{\tau_{j+1}} \mathbb{1}\{I_t = 1\}\right] && (86)\\
&\leqslant \sum_{j=\Gamma}^{T-1} \mathbb{E}\left[\frac{1-p^i_{1,\tau_j+1,T}}{p^i_{1,\tau_j+1,T}}\right], && (87)
\end{aligned}$$

where the inequality in Equation (87) uses the fact that $p^i_{1,t,T}$ is fixed, given $\mathcal{F}_{t-1}$. Then, we observe that $p^i_{1,t,T} = \mathbb{P}(\theta_{1,t,T} > y_i | \mathcal{F}_{t-1})$ changes only when the distribution of $\theta_{1,t,T}$ changes, that is, only on the time step after each play of the first arm. Thus, $p^i_{1,t,T}$ is the same at all time steps $t \in \{\tau_j+1, \ldots, \tau_{j+1}\}$, for every j. Finally, bounding the probability of selecting the optimal arm by 1 we have:

$$\text{(B)} \leqslant \sum_{j=\Gamma}^{T-1} \mathbb{E}\left[\frac{1}{p^i_{1,\tau_j+1,T}} - 1\right]. \tag{88}$$

By definition $N_{1,\tau_j+1} = j$, and let $S_{1,t} = s$. Thus, we have:

$$p^i_{1,\tau_j+1,T} = \mathbb{P}(\theta_{1,\tau_j+1,T} > y_i) = F^B_{j+1,y_i}(s)$$

due to the relation that links the Beta and the binomial distributions (Fact 3 of Agrawal and Goyal (2017)). Let $\tau_j + 1$ denote the time step after the j-th play of the optimal arm. Then, $N_{1,\tau_j+1} = j$. We do notice a sensible difference with respect to the stationary case. Indeed, the number of successes after j trial is not distributed anymore as a binomial distribution. Instead, it can be described by a Poisson-Binomial distribution $\mathrm{PB}(\underline{\mu}_1(j))$ where the vector $\underline{\mu}_1(j) = (\mu_1(1), \dots, \mu_1(j))$, and $\mu_1(m)$ represents the probability of success of the best arm at the m-th trial. The probability of having s successful trials out of a total of j trials can be written as follows (Wang, 1993; Poisson, 1837):

$$f_{j,\underline{\mu}_1(j)}(s) = \sum_{A \in F_s} \prod_{m \in A} \mu_1(m) \prod_{m' \in A^c} (1 - \mu_1(m')), \tag{89}$$

where F_s is the set of all subsets of s integers that can be selected from $\llbracket j \rrbracket$. F_s by definition will contain $\frac{j!}{(j-s)!s!}$ elements, the sum over which is infeasible to compute in practice unless the number of trials j is small. A useful property of f is that it is invariant to the order of the elements in $\underline{\mu}_1(j)$. Moreover, we define density function of the binomial of j trials and mean $\overline{\mu}_1(j)$, i.e., $\mathrm{Bin}(j, \overline{\mu}_1(j))$, as:

$$f_{j,\overline{\mu}_1(j)}(s) = \binom{j}{s} \overline{\mu}_1(j)^s (1 - \overline{\mu}_1(j))^{j-s}. \tag{90}$$

Using the first inequality of Lemma 5.2, we obtain:

$$\sum_{s=0}^{j} \frac{f_{j,\underline{\mu}_1(j)}(s)}{F^B_{j+1,y_i}(s)} \leqslant \sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(j)}(s)}{F^B_{j+1,y_i}(s)} \tag{91}$$

For the case $j \geqslant \sigma$, we apply the second inequality of Lemma 5.2, to obtain:

$$\sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(j)}(s)}{F^B_{j+1,y_i}(s)} \leqslant \sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(\sigma)}(s)}{F^B_{j+1,y_i}(s)}. \tag{92}$$

Instead, for the case $j < \sigma$, by applying the change of measure argument of Lemma C.2, we have:

$$\begin{aligned} \sum_{s=0}^{j} \frac{f_{j,\underline{\mu}_1(j)}(s)}{F^B_{j+1,y_i}(s)} \leqslant \sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(j)}(s)}{F^B_{j+1,y_i}(s)} &\leqslant \left(\frac{1}{(1-y_i)^{j+1}} - 1 \right) \delta_{\mathrm{TV}} \left(\mathrm{Bin}(j, \overline{\mu}_1(j)), \mathrm{Bin}(j, \overline{\mu}_1(\sigma)) \right) + \\ &\quad + \sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(\sigma)}(s)}{F^B_{j+1,y_i}(s)}, \end{aligned} \tag{93}$$

where $\delta_{TV}(P, Q) \coloneqq \sup_{A \in \mathcal{F}} \left| P(A) - Q(A) \right|$ is the total variation between the probability measures P and Q, having observed that, using the notation of Lemma C.2:

$$b = \max_{s \in \llbracket 0, j \rrbracket} \frac{1}{F^B_{j+1,y_i}(s)} = \frac{1}{F^B_{j+1,y_i}(0)} = \frac{1}{\mathbb{P}(\mathrm{Bin}(j+1, y_i) = 0)} = \frac{1}{(1-y_i)^{j+1}}, \tag{94}$$

$$a = \min_{s \in \llbracket 0, j \rrbracket} \frac{1}{F^B_{j+1,y_i}(s)} = \frac{1}{F^B_{j+1,y_i}(j)} = \frac{1}{\mathbb{P}(\mathrm{Bin}(j+1, y_i) \leqslant j)} \geqslant \frac{1}{\mathbb{P}(\mathrm{Bin}(j+1, y_i) \leqslant j+1)} = 1. \tag{95}$$

Putting all together, we obtain:

$$\begin{aligned} &\sum_{s=0}^{j} \frac{f_{j,\underline{\mu}_1(j)}(s)}{F^B_{j+1,y_i}(s)} \leqslant \\ &\leqslant \begin{cases} \left(\frac{1}{(1-y_i)^{j+1}} - 1 \right) \delta_{\mathrm{TV}}(\mathrm{Bin}(j, \overline{\mu}_1(j)), \mathrm{Bin}(j, \overline{\mu}_1(\sigma))) + \sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(\sigma)}(s)}{F^B_{j+1,y_i}(s)} & \text{if } 0 \leqslant j < \sigma \\ \sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(\sigma)}(s)}{F^B_{j+1,y_i}(s)} & \text{if } j \geqslant \sigma \end{cases}. \end{aligned} \tag{96}$$

Then, summing over j, we have:

$$\begin{aligned} &\sum_{j=\Gamma}^{T-1} \left(\sum_{s=0}^{j} \frac{f_{j,\underline{\mu}_1(j)}(s)}{F^B_{j+1,y_i}(s)} - 1 \right) \leqslant \\ &\leqslant \begin{cases} \sum_{j=\Gamma}^{\sigma} \left(\frac{1}{(1-y_i)^{j+1}} - 1 \right) \delta_{\mathrm{TV}}(\mathrm{Bin}(j, \overline{\mu}_1(j)), \mathrm{Bin}(j, \overline{\mu}_1(\sigma))) + \sum_{j=\Gamma}^{T-1} \left(\sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(\sigma)}(s)}{F^B_{j+1,y_i}(s)} - 1 \right) & \text{if } \Gamma < \sigma \\ \\ \sum_{j=\Gamma}^{T-1} \left(\sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(\sigma)}(s)}{F^B_{j+1,y_i}(s)} - 1 \right) & \text{if } \Gamma \geqslant \sigma \end{cases}, \end{aligned} \tag{97}$$

where we recall that $y_i \leqslant \overline{\mu}_1(\sigma)$. From Lemma 2.9 by Agrawal and Goyal (2017), we have that:

$$\sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(\sigma)}(s)}{F_{j+1,y_i}(s)} - 1 \leqslant \begin{cases} \frac{3}{\Delta'_i} & \text{if } j < \frac{8}{\Delta'_i} \\ \Theta\left(e^{-\frac{\Delta_i'^2 j}{2}} + \frac{e^{-D_i j}}{(j+1)\Delta_i'^2} + \frac{1}{e^{\Delta_i'^2 \frac{j}{4}}-1}\right) & \text{if } j \geqslant \frac{8}{\Delta'_i} \end{cases}, \tag{98}$$

where $\Delta'_i = \overline{\mu}_1(\sigma) - y_i$ and $D_i = d(y_i, \overline{\mu}_1(\sigma)) = y_i \log \frac{y_i}{\overline{\mu}_1(\sigma)} + (1-y_i) \log \frac{1-y_i}{1-\overline{\mu}_1(\sigma)}$. Thus, summing over j, we obtain:

$$\sum_{j=\Gamma}^{T-1} \left(\sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(\sigma)}(s)}{F_{j+1,y_i}(s)} - 1 \right) \leqslant \sum_{j<\min\{\frac{8}{\Delta'_i}-\Gamma,0\}} \frac{3}{\Delta'_i} + \sum_{j \geqslant \max\{\Gamma, \frac{8}{\Delta'_i}\}} \Theta\left(e^{-\frac{\Delta_i'^2 j}{2}} + \frac{e^{-D_i j}}{(j+1)\Delta_i'^2} + \frac{1}{e^{\Delta_i'^2 \frac{j}{4}}-1}\right) \tag{99}$$

$$\leqslant \sum_{j<\frac{8}{\Delta'_i}} \frac{3}{\Delta'_i} + \sum_{j \geqslant \frac{8}{\Delta'_i}} \Theta\left(e^{-\frac{\Delta_i'^2 j}{2}} + \frac{e^{-D_i j}}{(j+1)\Delta_i'^2} + \frac{1}{e^{\Delta_i'^2 \frac{j}{4}}-1}\right) \tag{100}$$

$$\leqslant \frac{24}{\Delta_i'^2} + \Theta\left(\frac{1}{\Delta_i'^2} + \frac{1}{\Delta_i'^2\sqrt{D_i}} + \frac{1}{\Delta_i'^3}\right), \tag{101}$$

having bounded the individual terms carefully:

$$\sum_{j \geqslant \frac{8}{\Delta'_i}} e^{-\frac{\Delta_i'^2 j}{2}} \leqslant \sum_{j=1}^{+\infty} e^{-\frac{\Delta_i'^2 j}{2}} \leqslant \int_{j=0}^{+\infty} e^{-\frac{\Delta_i'^2 j}{2}} \mathrm{d}j = \frac{2}{\Delta_i'^2}, \tag{102}$$

$$\sum_{j \geqslant \frac{8}{\Delta'_i}} \frac{e^{-D_i j}}{(j+1)\Delta_i'^2} \leqslant \frac{1}{8\Delta'_i} \sum_{j \geqslant \frac{8}{\Delta'_i}} e^{-D_i j} \leqslant \frac{1}{8\Delta'_i} \sum_{j=1}^{+\infty} e^{-D_i j} \leqslant \frac{1}{8\Delta'_i} \int_{j=0}^{+\infty} e^{-D_i j} \mathrm{d}j \leqslant \frac{1}{8\Delta'_i D_i}, \tag{103}$$

$$\sum_{j \geqslant \frac{8}{\Delta'_i}} \frac{1}{e^{\Delta_i'^2 \frac{j}{4}}-1} \leqslant \int_{j=\frac{4}{\Delta'_i}}^{+\infty} \frac{1}{e^{\Delta_i'^2 \frac{j}{4}}-1} \mathrm{d}j = \frac{4}{\Delta_i'^2} \log \frac{1}{1-e^{-\Delta'_i}} \leqslant \frac{4}{\Delta_i'^2} \log \frac{1}{\Delta'_i}, \tag{104}$$

having bounded $1 - e^{-x} \geqslant x$. First of all, let us observe that from Pinsker's inequality, we have $D_i \geqslant 2\Delta_i'^2$. We now relate $\Delta_i'^2$ with $d(\overline{\mu}_i(T), \overline{\mu}_1(\sigma))$. From the mean value theorem:

$$\frac{\epsilon}{(1+\epsilon)^2} d(\overline{\mu}_i(T), \overline{\mu}_1(\sigma)) = \frac{d(\overline{\mu}_i(T), \overline{\mu}_1(\sigma))}{1+\epsilon} - \frac{d(\overline{\mu}_i(T), \overline{\mu}_1(\sigma))}{(1+\epsilon)^2} \tag{105}$$

$$= d(x_i, \overline{\mu}_1(\sigma)) - d(x_i, y_i) \tag{106}$$

$$\leqslant \max_{z \in [y_i, \overline{\mu}_1(\sigma)]} \frac{\partial d(x_i, z)}{\partial z} (\overline{\mu}_1(\sigma) - y_i) \tag{107}$$

$$\leqslant \frac{\partial d(x_i, z)}{\partial z}\Big|_{z=\overline{\mu}_1(\sigma)} \Delta'_i, \tag{108}$$

having exploited that the maximum derivative is attained in $z = \overline{\mu}_1(\sigma)$. Moreover, we have, from Pinsker's inequality:

$$\frac{\partial d(x_i, z)}{\partial z}\Big|_{z=\overline{\mu}_1(\sigma)} = \frac{\overline{\mu}_1(\sigma) - x_i}{\overline{\mu}_1(\sigma)(1-\overline{\mu}_1(\sigma))} \leqslant \frac{\sqrt{d(x_i, \overline{\mu}_1(\sigma))}}{\sqrt{2}\overline{\mu}_1(\sigma)(1-\overline{\mu}_1(\sigma))} = \frac{\sqrt{d(\overline{\mu}_1(T), \overline{\mu}_1(\sigma))}}{\sqrt{2(1+\epsilon)}\overline{\mu}_1(\sigma)(1-\overline{\mu}_1(\sigma))}. \tag{109}$$

Putting all together, we obtain:

$$\Delta_i'^2 \geqslant \frac{2\epsilon^2}{(1+\epsilon)^3} (\overline{\mu}_1(\sigma)(1-\overline{\mu}_1(\sigma)))^2 d(\overline{\mu}_i(T), \overline{\mu}_1(\sigma)) = \frac{2\epsilon^2}{(1+\epsilon)^3} \widetilde{d}(\overline{\mu}_i(T), \overline{\mu}_1(\sigma)). \tag{110}$$

Finally, we get the following bound:

$$\sum_{j=\Gamma}^{T-1} \left(\sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(\sigma)}(s)}{F_{j+1,y_i}(s)} - 1 \right) \leqslant \Theta\left(\frac{1}{\epsilon^2 \widetilde{d}(\overline{\mu}_i(T), \overline{\mu}_1(\sigma))} \log \frac{1}{\epsilon^2 \widetilde{d}(\overline{\mu}_i(T), \overline{\mu}_1(\sigma))} \right). \tag{111}$$

Putting all together, we have:

$$\begin{aligned}\mathbb{E}[N_{i,T}] &\leqslant \Gamma + 2 + (1+\epsilon)^2 \frac{\log(T)}{d(\overline{\mu}_i(T), \overline{\mu}_1(\sigma))} \\ &+ \left(\log \frac{\overline{\mu}_1(\sigma)(1-\overline{\mu}_i(T))}{\overline{\mu}_i(T)(1-\overline{\mu}_1(\sigma))}\right)^2 \cdot \frac{(1+\epsilon)^2}{2\epsilon^2} \cdot \frac{1}{d(\overline{\mu}_i(T), \overline{\mu}_1(\sigma))^2} \\ &+ \Theta\left(\frac{1}{\epsilon^2 \widetilde{d}(\overline{\mu}_i(T), \overline{\mu}_1(\sigma))} \log \frac{1}{\epsilon^2 \widetilde{d}(\overline{\mu}_i(T), \overline{\mu}_1(\sigma))}\right) \\ &+ \sum_{j=\Gamma}^{\sigma} \frac{\delta_{\mathrm{TV}}(\mathrm{Bin}(j, \overline{\mu}_1(j)), \mathrm{Bin}(j, \overline{\mu}_1(\sigma))}{(1-\overline{\mu}_1(\sigma))^{j+1}}.\end{aligned} \tag{112}$$

By setting $\epsilon' = 3\epsilon$, we get the result. ☐

Theorem 6.3 (γ-`ET-GTS` Bound)**.** *Let $\sigma \in [\![\sigma_{\boldsymbol{\mu}}(T), T]\!]$. Under Assumptions 3.1 and 3.3, setting $\gamma \leqslant \min\left\{\frac{1}{4\lambda^2}, 1\right\}$, for the γ-`ET-GTS` algorithm, for every arm $i \in [\![K]\!] \backslash \{i^*(T)\}$, it holds that:*

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mu}}[N_{i,T}] \leqslant O\Bigg(&\underbrace{\Gamma}_{(i)} + \underbrace{\frac{\log(T\overline{\Delta}_i(\sigma,T)^2 + e^6)}{\gamma \overline{\Delta}_i(\sigma,T)^2}}_{(ii)} \\ &+ \underbrace{\sum_{j=\Gamma}^{\sigma} \frac{\delta_{\mathrm{TV}}(\mathbb{P}_j, \mathbb{Q}_j(\overline{\mu}_{i^*(T)}(\sigma)))}{\mathrm{erfc}(\sqrt{\gamma j/2}(\overline{\mu}_{i^*(T)}(\sigma)))}}_{(iii)}\Bigg),\end{aligned}$$

where $\mathrm{erfc}(\cdot)$ is the complementary error function, $\mathbb{P}_j$ is the distribution of the sample mean of the first j samples collected from arm $i^*(T)$, while $\mathbb{Q}_j(y)$ is the distribution of the sample mean of j samples collected from any *λ^2-subgaussian distribution with mean y.*

Proof. Using Lemma B.1 to bound the expected value of pulls of the suboptimal arm when the window length is equal to the time horizon we have:

$$\begin{aligned}\mathbb{E}_\tau[N_{i,T}] \leqslant \Gamma + \frac{1}{\epsilon_i} + \frac{T}{T} + \frac{\omega T}{T} + &\underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+1}^{T} \mathbb{1}\left\{p^i_{i,t,T} > \frac{1}{T\epsilon_i}, I_t = i, N_{i,t} \geqslant \omega\right\}\right]}_{\text{(A)}} \\ &+ \underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+1}^{T} \left(\frac{1}{p^i_{i^*(t),t,T}} - 1\right) \mathbb{1}\left\{I_t = i^*(t)\right\}\right]}_{\text{(B)}}.\end{aligned}$$

In the following, we consider, without loss of generality, the first arm to be the best, i.e., for each round $t \in [\![T]\!]$ we will have $i^*(t) = i^*(T) = 1$. Furthermore, we set $y_{i,t} = y_i = \overline{\mu}_1(\sigma) - \frac{\overline{\Delta}_i(\sigma,T)}{3}$, $x_{i,t} = x_i = \overline{\mu}_i(T) + \frac{\overline{\Delta}_i(\sigma,T)}{3}$, $\omega = \frac{32 \log(T\overline{\Delta}_i(\sigma,T)^2 + e^6)}{\gamma(y_i - x_i)^2} = \frac{288 \log(T\overline{\Delta}_i(\sigma,T)^2 + e^6)}{\gamma \overline{\Delta}_i(\sigma,T)^2}$, $\epsilon_i = \overline{\Delta}_i(\sigma,T)^2$, $l = 0$. Finally, we define $\hat{\overline{\mu}}_{i,t,T} = \frac{S_{i,t}}{N_{i,t}}$. We rewrite term (A) as a contribution of two different terms.

$$\begin{aligned}\text{(A)} = &\underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+1}^{T} \mathbb{1}\left\{p^i_{i,t,T} > \frac{1}{T\epsilon_i}, \hat{\overline{\mu}}_{i,t,T} \leqslant x_i, I_t = i, N_{i,t} \geqslant \omega\right\}\right]}_{\text{(A1)}} \\ &+ \underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+1}^{T} \mathbb{1}\left\{p^i_{i,t,T} > \frac{1}{T\epsilon_i}, \hat{\overline{\mu}}_{i,t,T} > x_i, I_t = i, N_{i,t} \geqslant \omega\right\}\right]}_{\text{(A2)}}.\end{aligned} \tag{113}$$

We focus first on term (A2).

Term (A2) Focusing on (A2), let τ_j^i denote the round in which arm i is played for the j-th time:

$$\text{(A2)} \leqslant \mathbb{E}\left[\sum_{t=K\Gamma+1}^{T} \mathbb{1}\left\{\overline{\hat{\mu}}_{i,t,T} > x_i, I_t = i, N_{i,t} \geqslant \omega\right\}\right] \tag{114}$$

$$\leqslant \mathbb{E}\left[\sum_{j=\omega}^{T} \mathbb{1}\left\{\overline{\hat{\mu}}_{i,\tau_j^i+1,T} > x_i\right\} \sum_{t=\tau_j^i+1}^{\tau_{j+1}^i} \mathbb{1}\left\{I_t = i\right\}\right] \tag{115}$$

$$\leqslant \sum_{j=\omega}^{T} \mathbb{P}(\overline{\hat{\mu}}_{i,\tau_j^i+1,T} > x_i) \tag{116}$$

$$= \sum_{j=\omega}^{T} \mathbb{P}(\overline{\hat{\mu}}_{i,\tau_j^i+1,T} - \mathbb{E}[\overline{\hat{\mu}}_{i,\tau_j^i+1,T}] > x_i - \mathbb{E}[\overline{\hat{\mu}}_{i,\tau_j^i+1,T}]) \tag{117}$$

$$\tag{118}$$

$$\leqslant \sum_{j=\omega}^{T} \mathbb{P}(\overline{\hat{\mu}}_{i,\tau_j^i+1,T} - \mathbb{E}[\overline{\hat{\mu}}_{i,\tau_j^i+1,T}] > x_i - \overline{\mu}_i(T)) \tag{119}$$

$$\leqslant \sum_{j=\omega}^{T} \exp\left(-\frac{1}{2\lambda^2}\omega(x_i - \overline{\mu}_i(T))^2\right), \tag{120}$$

where the last inequality follows from the Chernoff-Hoeffding bound (Lemma C.9). By definition, we will have that $x_i - \overline{\mu}_i(T) = y_i - x_i$. So, substituting in the last inequality, we obtain:

$$\sum_{j=\omega}^{T} \exp\left(-\frac{1}{2\lambda^2}\omega(x_i - \overline{\mu}_i(T))^2\right) = \sum_{j=\omega}^{T} \exp\left(-2\frac{32\log(T\overline{\Delta}_i(\sigma,T)^2 + e^6)}{(y_i - x_i)^2}(x_i - \overline{\mu}_i(T))^2\right) \tag{121}$$

$$\leqslant \frac{1}{\overline{\Delta}_i(\sigma,T)^2}. \tag{122}$$

Term (A1) Focusing on term (A1), we have:

$$\text{(A1)} \leqslant \mathbb{E}\left[\sum_{t=K\Gamma+1}^{T} \mathbb{1}\left\{\underbrace{p_{i,t,T}^i > \frac{1}{T\epsilon_i}, N_{i,t} \geqslant \omega, \overline{\hat{\mu}}_{i,t,T} \leqslant x_i}_{\text{(C)}}\right\}\right]. \tag{123}$$

We now wish to evaluate if condition (C) ever occurs. In order to do so, notice that $\theta_{i,t,T}$ is a normal distributed random variable, in particular distributed as $\mathcal{N}\left(\overline{\hat{\mu}}_{i,t,T}, \frac{1}{\gamma N_{i,t}}\right)$. An $\mathcal{N}\left(m, \sigma^2\right)$ distributed random variable is stochastically dominated by $\mathcal{N}\left(m', \sigma^2\right)$ distributed r.v. if $m' \geqslant m$. Therefore, given $\overline{\hat{\mu}}_{i,t,T} \leqslant x_i$, the distribution of $\theta_{i,t,T}$ is stochastically dominated by $\mathcal{N}\left(x_i, \frac{1}{\gamma N_{i,t}}\right)$ (see Lemma C.17). This implies:

$$\mathbb{P}\left(\theta_{i,t,T} > y_i \mid N_{i,t} \geqslant \omega, \overline{\hat{\mu}}_{i,t,T} \leqslant x_i, \mathbb{F}_{t-1}\right) \leqslant \mathbb{P}\left(\mathcal{N}\left(x_i, \frac{1}{\gamma N_{i,t}}\right) > y_i \middle| \mathcal{F}_{t-1}, N_{i,t} > \omega\right).$$

Using Lemma C.8, we have:

$$\mathbb{P}\left(\mathcal{N}\left(x_i, \frac{1}{\gamma N_{i,t}}\right) > y_i\right) \leqslant \frac{1}{2}e^{-\frac{(\gamma N_{i,t})(y_i - x_i)^2}{2}} \tag{124}$$

$$\leqslant \frac{1}{2}e^{-\frac{(\gamma\omega)(y_i - x_i)^2}{2}}, \tag{125}$$

which is smaller than $\frac{1}{T\overline{\Delta}_i(\sigma,T)^2}$ because $\omega \geqslant \frac{2\log\left(T\overline{\Delta}_i(\sigma,T)^2\right)}{\gamma(y_i - x_i)^2}$. Substituting, we get,

$$\mathbb{P}\left(\theta_{i,t,T} > y_i \mid N_{i,t} > \omega, \hat{\mu}_i(t) \leqslant x_i, \mathcal{F}_{t-1}\right) \leqslant \frac{1}{T\overline{\Delta}_i(\sigma,T)^2}. \tag{126}$$

In contradiction with the condition (C), then (A1) $= 0$.

Term (B) Focusing on term (B):

$$(\mathrm{B}) = \mathbb{E}\left[\sum_{t=K\Gamma+1}^{T}\left(\frac{1}{p^i_{1,t,T}}-1\right)\mathbb{1}\left\{I_t=1\right\}\right] \tag{127}$$

Let τ_j denote the time step at which arm 1 is played for the j-th time:

$$(\mathrm{B}) \leqslant \sum_{j=\Gamma}^{T-1}\mathbb{E}\Bigg[\frac{1-p^i_{1,\tau_j+1,T}}{p^i_{1,\tau_j+1,T}}\sum_{t=\tau_j+1}^{\tau_{j+1}}\mathbb{1}\{I_t=1\}\Bigg] \tag{128}$$

$$\leqslant \sum_{j=\Gamma}^{T-1}\mathbb{E}\Bigg[\frac{1-p^i_{1,\tau_j+1,T}}{p^i_{1,\tau_j+1,T}}\Bigg], \tag{129}$$

where the inequality in Equation (129) uses the fact that $p^i_{1,t,T}$ is fixed, given $\mathcal{F}_{t-1}$. Then, we observe that $p^i_{1,t,T}=\mathbb{P}(\theta_{1,t,T}>y_i|\mathcal{F}_{t-1})$ changes only when the distribution of $\theta_{1,t,T}$ changes, that is, only on the time step after each play of the first arm. Thus, $p^i_{1,t,T}$ is the same at all time steps $t\in\{\tau_j+1,\dots,\tau_{j+1}\}$, for every j. Finally, bounding the probability of selecting the optimal arm by 1 we have:

$$(\mathrm{B}) \leqslant \sum_{j=\Gamma}^{T-1}\mathbb{E}\left[\frac{1}{p^i_{1,\tau_j+1,T}}-1\right]. \tag{130}$$

Now in order to face this term, let us consider the arbitrary trial j. Thanks to Lemma C.2, we can bound the difference between the real process and an analogous (same number of trials) virtual process with mean $\overline{\mu}_1(\sigma)$ (where by stationary we mean that all the trials of the virtual process will have a fixed mean):

$$\underbrace{\mathbb{E}\left[\frac{1}{p^i_{1,\tau_j+1,T}}\right]}_{(\mathrm{D})} \leqslant \frac{2\delta_{TV}(\mathbb{P}_j,\mathbb{Q}_j(\overline{\mu}_1(\sigma)))}{\mathrm{erfc}(\sqrt{\frac{\gamma j}{2}}\overline{\mu}_1(\sigma))}+\underbrace{\mathbb{E}_{\overline{\mu}_1(\sigma)}\left[\frac{1}{p^i_{1,\tau_j+1,T}}\right]}_{(\mathrm{E})}, \tag{131}$$

where $\mathbb{P}_j$ is the distribution of the sample mean of the first j samples collected from arm 1, namely $\overline{\hat{\mu}}_{1,\tau_j+1,T}$, while $\mathbb{Q}_j(y)$ is the distribution of the sample mean of j samples collected from *any* λ^2-subgaussian distribution with fixed mean $\overline{\mu}_1(\sigma)$. By definition, $\delta_{TV}(P,Q) := \sup_{A\in\mathcal{F}}\left|P(A)-Q(A)\right|$ is the total variation between the probability measures P and Q (assuming they are defined over a measurable space $(\Omega,\mathcal{F})$), having observed that, using the notation of Lemma C.2 (as the environment cannot generate rewards smaller than zero):

$$b = \max_{s\geqslant 0}\frac{1}{\mathbb{P}\left(\mathcal{N}\left(s,\frac{1}{\gamma N_{i,t}}\right)>y_i\right)} \leqslant \frac{1}{\mathbb{P}\left(\mathcal{N}\left(0,\frac{1}{\gamma N_{i,t}}\right)\geqslant\overline{\mu}_1(\sigma)\right)} = \frac{2}{\mathrm{erfc}(\sqrt{\frac{\gamma j}{2}}\overline{\mu}_1(\sigma))}, \tag{132}$$

$$a = 1. \tag{133}$$

Our interest is to find if there is a minimum number of trials j such that we will have (D) $\geqslant$ (E) without the need to add any term. Given $\mathcal{F}_{\tau_j}$, let Θ_j denote a $\mathcal{N}\left(\overline{\hat{\mu}}_1\left(\tau_j+1\right),\frac{1}{\gamma j}\right)$ distributed Gaussian random variable. Let G_j be the geometric random variable denoting the number of consecutive independent trials until and including the trial where a sample of Θ_j becomes greater than y_i. Then observe that $p^i_{1,\tau_j+1,T}=\Pr\left(\Theta_j>y_i\mid\mathcal{F}_{\tau_j}\right)$ and

$$\mathbb{E}\left[\frac{1}{p^i_{1,\tau_j+1,T}}\right] = \mathbb{E}\left[\mathbb{E}\left[G_j\mid\mathcal{F}_{\tau_j}\right]\right] = \mathbb{E}\left[G_j\right]. \tag{134}$$

We compute first the expected value for the real process. We will consider first j such that $\overline{\mu}_1(j)\geqslant\overline{\mu}_1(\sigma)$, we will bound the expected value of G_j by a constant for all j defined as earlier. Consider any integer $r\geqslant 1$. Let $z=\sqrt{\log r}$ and let random variable MAX $_r$ denote the maximum of r independent samples of Θ_j. We abbreviate $\overline{\hat{\mu}}_1\left(\tau_j+1\right)$ to $\overline{\hat{\mu}}_1$ and we will abbreviate $\overline{\mu}_1(\sigma)$ as μ_1 and $\overline{\Delta}_i(\sigma,T)$ as Δ_i in the following. Then for any integer $r\geqslant 1$:

$$\mathbb{P}\left(G_j\leqslant r\right) \geqslant \mathbb{P}\left(\mathrm{MAX}_r>y_i\right) \tag{135}$$

$$\geqslant \mathbb{P}\left(\mathrm{MAX}_r>\overline{\hat{\mu}}_1+\frac{z}{\sqrt{\gamma j}}\geqslant y_i\right) \tag{136}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(\mathrm{MAX}_r>\overline{\hat{\mu}}_1+\frac{z}{\sqrt{\gamma j}}\geqslant y_i\right)\middle|\ \mathcal{F}_{\tau_j}\right]\right] \tag{137}$$

$$= \mathbb{E}\left[\mathbb{1}\left(\overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant y_i\right) \mathbb{P}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \,\middle|\, \mathcal{F}_{\tau_j}\right)\right]. \tag{138}$$

For any instantiation F_{τ_j} of $\mathcal{F}_{\tau_j}$, since Θ_j is Gaussian $\mathcal{N}\left(\hat{\mu}_1, \frac{1}{\gamma j}\right)$ distributed r.v., this gives using C.7:

$$\mathbb{P}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \,\middle|\, \mathcal{F}_{\tau_j} = F_{\tau_j}\right) \geqslant 1 - \left(1 - \frac{1}{\sqrt{2\pi}} \frac{z}{(z^2+1)} e^{-z^2/2}\right)^r \tag{139}$$

$$= 1 - \left(1 - \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\log r}}{(\log r + 1)} \frac{1}{\sqrt{r}}\right)^r \tag{140}$$

$$\geqslant 1 - e^{-\frac{r}{\sqrt{4\pi r \log r}}}. \tag{141}$$

For $r \geqslant e^{12}$:

$$\mathbb{P}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \,\middle|\, \mathcal{F}_{\tau_j} = F_{\tau_j}\right) \geqslant 1 - \frac{1}{r^2}. \tag{142}$$

Substituting we obtain:

$$\mathbb{P}\left(G_j \leqslant r\right) \geqslant \mathbb{E}\left[\mathbb{1}\left(\overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant y_i\right)\left(1 - \frac{1}{r^2}\right)\right] \tag{143}$$

$$= \left(1 - \frac{1}{r^2}\right) \mathbb{P}\left(\overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant y_i\right). \tag{144}$$

Applying Lemma C.9 to the second term, we can write, since $\mathbb{E}[\overline{\hat{\mu}}_1] \geqslant \mu_1$:

$$\mathbb{P}\left(\overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant \mu_1\right) \geqslant 1 - e^{-\frac{z^2}{2\gamma\lambda^2}} \geqslant 1 - \frac{1}{r^2}, \tag{145}$$

being $\gamma \leqslant \frac{1}{4\lambda^2}$. Using, $y_i \leqslant \mu_1$, this gives

$$\mathbb{P}\left(\overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant y_i\right) \geqslant 1 - \frac{1}{r^2}. \tag{146}$$

Substituting all back we obtain:

$$\mathbb{E}\left[G_j\right] = \sum_{r=0}^{+\infty} \mathbb{P}\left(G_j \geqslant r\right) \tag{147}$$

$$= 1 + \sum_{r=1}^{\infty} \mathbb{P}\left(G_j \geqslant r\right) \tag{148}$$

$$\leqslant 1 + e^{12} + \sum_{r \geqslant 1}\left(\frac{1}{r^2} + \frac{1}{r^2}\right) \tag{149}$$

$$\leqslant 1 + e^{12} + 2 + 2. \tag{150}$$

This shows a constant bound of $\mathbb{E}\left[\frac{1}{p^i_{1,\tau_j+1,T}} - 1\right] = \mathbb{E}\left[G_j\right] - 1 \leqslant e^{12} + 5$ for all $j \geqslant \sigma$. We derive a bound for large j. Consider $j \geqslant \omega$ (and still $j \geqslant \sigma$). Given any $r \geqslant 1$, define G_j, MAX_r, and $z = \sqrt{\log r}$ as defined earlier. Then,

$$\mathbb{P}\left(G_j \leqslant r\right) \geqslant \mathbb{P}\left(\mathrm{MAX}_r > y_i\right) \tag{151}$$

$$\geqslant \mathbb{P}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6} \geqslant y_i\right) \tag{152}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6} \geqslant y_i\right) \,\middle|\, \mathcal{F}_{\tau_j}\right]\right] \tag{153}$$

$$= \mathbb{E}\left[\mathbb{1}\left(\overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} + \frac{\Delta_i}{6} \geqslant \mu_1\right) \mathbb{P}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6} \,\middle|\, \mathcal{F}_{\tau_j}\right)\right]. \tag{154}$$

where we used that $y_i = \mu_1 - \frac{\Delta_i}{3}$. Now, since $j \geqslant \omega = \frac{288 \log\left(T\Delta_i^2 + e^6\right)}{\gamma \Delta_i^2}$,

$$2\frac{\sqrt{2\log\left(T\Delta_i^2 + e^6\right)}}{\sqrt{\gamma j}} \leqslant \frac{\Delta_i}{6}. \tag{155}$$

Therefore, for $r \leqslant \left(T\Delta_i^2 + e^6\right)^2$,

$$\frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6} = \frac{\sqrt{\log(r)}}{\sqrt{\gamma j}} - \frac{\Delta_i}{6} \leqslant -\frac{\Delta_i}{12}. \tag{156}$$

Then, since Θ_j is $\mathcal{N}\left(\overline{\hat{\mu}}_1\left(\tau_j+1\right), \frac{1}{\gamma j}\right)$ distributed random variable, using the upper bound in Lemma C.8, we obtain for any instantiation F_{τ_j} of history $\mathbb{F}_{\tau_j}$,

$$\mathbb{P}\left(\Theta_j > \overline{\hat{\mu}}_1\left(\tau_j+1\right) - \frac{\Delta_i}{12} \middle| \mathcal{F}_{\tau_j} = F_{\tau_j}\right) \geqslant 1 - \frac{1}{2} e^{-\gamma j \frac{\Delta_i^2}{288}} \geqslant 1 - \frac{1}{2\left(T\Delta_i^2 + e^6\right)}. \tag{157}$$

being $j \geqslant \omega$. This implies:

$$\mathbb{P}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1\left(\tau_j+1\right) + \frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6} \middle| \mathcal{F}_{\tau_j} = F_{\tau_j}\right) \geqslant 1 - \frac{1}{2^r\left(T\Delta_i^2 + e^6\right)^r}. \tag{158}$$

Also, for any $t \geqslant \tau_j + 1$, using Lemma C.9, as $\mathbb{E}[\overline{\hat{\mu}}_1] \geqslant \mu_1$ we get:

$$\mathbb{P}\left(\overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6} \geqslant y_i\right) \geqslant \mathbb{P}\left(\overline{\hat{\mu}}_1 \geqslant \mu_1 - \frac{\Delta_i}{6}\right) \geqslant 1 - e^{-j\Delta_i^2/72\lambda^2} \geqslant 1 - \frac{1}{\left(T\Delta_i^2 + e^6\right)^{16}}. \tag{159}$$

Let $T' = \left(T\Delta_i^2 + e^6\right)^2$. Therefore, for $1 \leqslant r \leqslant T'$, we have:

$$\mathbb{P}\left(G_j \leqslant r\right) \geqslant 1 - \frac{1}{2^r\left(T'\right)^{r/2}} - \frac{1}{\left(T'\right)^8}. \tag{160}$$

When $r \geqslant T' \geqslant e^{12}$, we obtain:

$$\mathbb{P}\left(G_j \leqslant r\right) \geqslant 1 - \frac{1}{r^2} - \frac{1}{r^2}. \tag{161}$$

Combining all the bounds we have derived:

$$\begin{aligned}
\mathbb{E}\left[G_j\right] &\leqslant \sum_{r=0}^{\infty} \mathbb{P}\left(G_j \geqslant r\right) && (162)\\
&\leqslant 1 + \sum_{r=1}^{T'} \mathbb{P}\left(G_j \geqslant r\right) + \sum_{r=T'}^{\infty} \mathbb{P}\left(G_j \geqslant r\right) && (163)\\
&\leqslant 1 + \sum_{r=1}^{T'} \frac{1}{\left(2\sqrt{T'}\right)^r} + \frac{1}{\left(T'\right)^7} + \sum_{r=T'}^{\infty} \frac{1}{r^2} + \frac{1}{r^{1.5}} && (164)\\
&\leqslant 1 + \frac{1}{\sqrt{T'}} + \frac{1}{\left(T'\right)^7} + \frac{2}{T'} + \frac{3}{\sqrt{T'}} && (165)\\
&\leqslant 1 + \frac{5}{T\Delta_i^2 + e^6}. && (166)
\end{aligned}$$

So we have proved that:

$$\mathbb{E}\left[\frac{1}{p^i_{1,\tau_j+1,T}} - 1\right] \leqslant \begin{cases} \frac{2\delta_{TV}\left(\mathbb{P}_j\left(\overline{\mu}_1(j)\right), \mathbb{Q}_j\left(\overline{\mu}_1(\sigma)\right)\right)}{\mathrm{erfc}\left(\sqrt{\frac{\gamma j}{2}}\,\overline{\mu}_1(\sigma)\right)} + \mathbb{E}_{\overline{\mu}_1(\sigma)}\left[\frac{1}{p^i_{1,\tau_j+1,T}} - 1\right]. & \text{if } 0 \leqslant j < \sigma \\ \left(e^{12} + 5\right) & \text{if } j \geqslant \sigma \\ \frac{5}{T\overline{\Delta}_i(\sigma,T)^2} & \textit{if } j \geqslant \omega \textit{ and } j \geqslant \sigma \end{cases} \tag{167}$$

Furthermore, it also holds:

$$\mathbb{E}_{\overline{\mu}_1(\sigma)}\left[\frac{1}{p^i_{1,\tau_j+1,T}} - 1\right] \leqslant \begin{cases} \left(e^{12} + 5\right) & \text{if } j \leqslant \omega \\ \frac{5}{T\overline{\Delta}_i(\sigma,T)^2} & \text{if } j \geqslant \omega \end{cases}, \tag{168}$$

as it is the expected value for the term in a stationary process with fixed mean for the reward equal to $\overline{\mu}_1(\sigma)$ (Lemma 6 Agrawal and Goyal (2013), but can be also retrieved by making the same calculation that we have just performed). So, we can write:

$$
\begin{aligned}
\text{(B)} &\leqslant \sum_{j=\Gamma}^{T-1} \mathbb{E}\left[\frac{1}{p^i_{1,\tau_j+1,T}} - 1\right] \\
&\leqslant \begin{cases} \sum_{j=\Gamma}^{\omega}(e^{12}+5) + \sum_{j=\Gamma}^{T-1} \frac{1}{T\overline{\Delta}_i(\sigma,T)^2} & \text{if } \Gamma \geqslant \sigma \\ \\ \sum_{j=\Gamma}^{\sigma} \frac{2\delta_{TV}(\mathbb{P}_j(\overline{\mu}_1(j)),\mathbb{Q}_j(\overline{\mu}_1(\sigma)))}{\mathrm{erfc}(\sqrt{\frac{\gamma j}{2}}\overline{\mu}_1(\sigma))} + \sum_{j=\Gamma}^{\omega}(e^{12}+5) + \sum_{j=\Gamma}^{T-1} \frac{1}{T\overline{\Delta}_i(\sigma,T)^2} & \text{if } \Gamma < \sigma \end{cases} && (169) \\
&\leqslant \begin{cases} \omega(e^{12}+5) + \frac{1}{\overline{\Delta}_i(\sigma,T)^2} & \text{if } \Gamma \geqslant \sigma \\ \\ \sum_{j=\Gamma}^{\sigma} \frac{2\delta_{TV}(\mathbb{P}_j(\overline{\mu}_1(j)),\mathbb{Q}_j(\overline{\mu}_1(\sigma)))}{\mathrm{erfc}(\sqrt{\frac{\gamma j}{2}}\overline{\mu}_1(\sigma))} + \omega(e^{12}+5) + \frac{1}{\overline{\Delta}_i(\sigma,T)^2} & \text{if } \Gamma < \sigma \end{cases}, && (170)
\end{aligned}
$$

summing all the term follows the statement. ☐

Corollary 6.4 (Explicit `Beta-TS` and γ-`GTS` Bound)**.** *Let $\overline{\sigma} \geqslant 0$. Under the same assumptions of Theorems 6.2 and 6.3, setting $\Gamma = \alpha\overline{\sigma}$, with $\alpha \geqslant 1$, for both* `ET-Beta-TS` *and γ-*`ET-GTS`*, for every $\boldsymbol{\mu} \in \mathcal{M}_{\overline{\sigma}}$ and for every arm $i \in [\![K]\!]\backslash\{i^*(T)\}$, it holds that:*[17]

$$
\mathbb{E}_{\boldsymbol{\mu}}[N_{i,T}] \leqslant O\Bigg(\underbrace{\overline{\sigma}}_{(i)} + \underbrace{\frac{\log(T)}{\overline{\Delta}_i(\overline{\sigma},T)^2}}_{(ii)}\Bigg).
$$

Proof. The proof for `Beta-TS` follows from the proof of Theorem 6.2, setting $\omega = \frac{\log(T)}{2(x_i-y_i)^2}$, $x_i = \overline{\mu}_i(T) + \frac{\overline{\Delta}_i(\sigma,T)}{3}$ and $y_i = \overline{\mu}_1(\sigma) - \frac{\overline{\Delta}_i(\sigma,T)}{3}$. Rewriting Equation (72):

$$
\sum_{j=0}^{T} \exp(-jd(x_i,y_i)) \leqslant 1 + \sum_{j=1}^{T} \frac{9}{j(x_i-y_i)^2} \leqslant 1 + \frac{9\ln(T)}{\overline{\Delta}_i(\sigma,T)^2}. \tag{171}
$$

Finally we can rewrite Equation (98):

$$
\begin{aligned}
\sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(\sigma)}(s)}{F_{j+1,y_i}(s)} - 1 &\leqslant \begin{cases} \frac{3}{\Delta'_i} & \text{if } j < \frac{8}{\Delta'_i} \\ \Theta\left(e^{-\frac{\Delta_i'^2 j}{2}} + \frac{e^{-D_i j}}{(j+1)\Delta_i'^2} + \frac{1}{e^{\Delta_i'^2 \frac{j}{4}}-1}\right) & \text{if } j \geqslant \frac{8}{\Delta'_i} \end{cases}, && (172) \\
&\leqslant \begin{cases} \frac{3}{\Delta'_i} & \text{if } j < \frac{8}{\Delta'_i} \\ \Theta\left(\frac{2}{\Delta_i'^2 j} + \frac{1}{(j+1)\Delta_i'^2} + \frac{1}{\Delta_i'^2 \frac{j}{4}}\right) & \text{if } j \geqslant \frac{8}{\Delta'_i} \end{cases} && (173)
\end{aligned}
$$

where $\Delta'_i = \overline{\mu}_1(\sigma) - y_i = \frac{\overline{\Delta}_i(\sigma,T)}{3}$ and $D_i = d(y_i, \overline{\mu}_1(\sigma)) = y_i \log \frac{y_i}{\overline{\mu}_1(\sigma)} + (1-y_i)\log\frac{1-y_i}{1-\overline{\mu}_1(\sigma)}$, where the inequalities follow from the facts that $e^{-x} \leqslant \frac{1}{x}$ (for $x \geqslant 0$) and $e^x \geqslant 1 + x$ (for every value of x). Summing over all the terms we obtain:

$$
\sum_{j=1}^{T}\left(\sum_{s=0}^{j} \frac{f_{j,\overline{\mu}_1(\sigma)}(s)}{F_{j+1,y_i}(s)} - 1\right) \leqslant O\left(\frac{1}{\overline{\Delta}_i(\sigma,T)^2} + \frac{\log(T)}{\overline{\Delta}_i(\sigma,T)^2}\right). \tag{174}
$$

By doing all the steps from Equation (79) to Equation (83) it is easy to see that the condition from Equation (84) still holds. The result follows by summing all the terms and noticing that for all the instances in $\mathcal{M}_\sigma$, whenever each arm is pulled at least σ times the sum of the dissimilarity terms vanishes. The result for γ-`GTS` follows trivially from the statement of Theorem 6.3, noting again that by definition of the class of instances all the dissimilarity terms vanish for $\Gamma \geqslant \sigma$. ☐

Theorem 6.5 (Lower Bound)**.** *Let $T \in \mathbb{N}$ and $\overline{\sigma} \in [\![2, \frac{T-1}{2}]\!]$. For every algorithm $\mathfrak{A}$, it holds that:*

$$
\sup_{\boldsymbol{\mu} \in \mathcal{M}^{\text{det}}_{\overline{\sigma}}} R_{\boldsymbol{\mu}}(\mathfrak{A}, T) \geqslant \frac{K}{64}(\overline{\sigma} - 2). \tag{6}
$$

---

[17] Notice that the choice of the exploration parameter Γ requires just an upper bound $\overline{\sigma}$ to the complexity index $\sigma_{\boldsymbol{\mu}}(T)$ of the set of SRRB of interest $\mathcal{M}_{\overline{\sigma}}$. Details are provided in Appendix D.

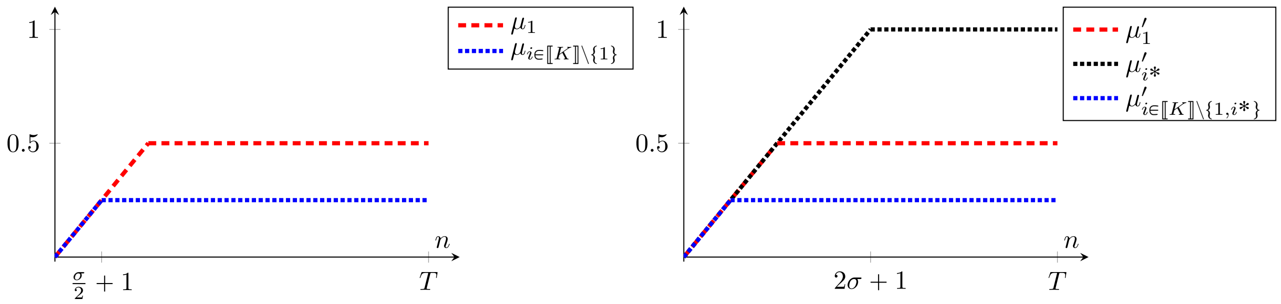


Figure 7: The two instances $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$.

Proof. First of all, we build two instances, defined for $n \in [\![T]\!]$ and $2\sigma + 1 \leqslant T \implies \sigma \leqslant (T-1)/2$:

$$\boldsymbol{\mu} = \begin{cases} \mu_1(n) = \min\left\{\frac{n-1}{2\sigma}, \frac{1}{2}\right\} \\ \\ \mu_i(n) = \min\left\{\frac{n-1}{2\sigma}, \frac{1}{4}\right\} \quad \text{if } i \in [\![K]\!]\setminus\{1\} \end{cases}, \tag{175}$$

We define $\tau_{i,\sigma}$ as the round in which the i-th arm gets to $\frac{\sigma}{2}$ pulls and $i^* \in \arg\max_{i\in[\![K]\!]\setminus\{1\}} \{\mathbb{E}_{\boldsymbol{\mu}}[\tau_{i,\sigma}]\}$, i.e., the arm whose expected value for the round in which it is played for the $\frac{\sigma}{2}$-th time under policy $\mathfrak{A}$ in environment $\boldsymbol{\mu}$ is the largest in the set of the suboptimal arms. We then introduce the second environment as:

$$\boldsymbol{\mu}' = \begin{cases} \mu'_1(n) = \min\left\{\frac{n-1}{2\sigma}, \frac{1}{2}\right\} \\ \\ \mu'_i(n) = \min\left\{\frac{n-1}{2\sigma}, \frac{1}{4}\right\} \quad \text{if } i \in [\![K]\!]\setminus\{1, i^*\} \\ \\ \mu'_{i^*}(n) = \min\left\{\frac{n-1}{2\sigma}, 1\right\} \end{cases}. \tag{176}$$

Intuitively, the two environments are indistinguishable as long as an algorithm does not pull arm i^* at least $\frac{\sigma}{2}$ times. The instances are depicted in Figure 7.

By performing some simple calculations, the following equalities also hold:

$$\begin{aligned} &\overline{\mu}_i(T) = \overline{\mu}_{i^*}(T) = \overline{\mu}'_i(T) = \frac{1}{4} - \frac{\frac{\sigma}{2}+1}{8T}, \\ &\overline{\mu}'_{i^*}(T) = 1 - \frac{2\sigma+1}{2T}, && \text{if } i \notin [\![K]\!]\setminus\{1, i^*\} \\ &\overline{\mu}_1(T) = \overline{\mu}'_1(T) = \frac{1}{2} - \frac{\sigma+1}{4T} \end{aligned}$$

Consequently, we have for all $i \in [\![K]\!]\setminus\{1\}$:

$$\overline{\Delta}_{i,\boldsymbol{\mu}}(T,T) = \overline{\mu}_1(T) - \overline{\mu}_i(T) = \frac{1}{4} - \frac{\frac{3\sigma}{2}+1}{8T} \geqslant \frac{\frac{5}{4}T+1}{8T} \geqslant \frac{5}{32}, \tag{177}$$

and again for all $i \in [\![K]\!]\setminus\{i^*\}$:

$$\overline{\Delta}_{i,\boldsymbol{\mu}'}(T,T) \geqslant \overline{\Delta}_{1,\boldsymbol{\mu}'}(T,T) \geqslant \overline{\mu}'_{i^*}(T) - \overline{\mu}'_1(T) = \frac{1}{2} - \frac{3\sigma+1}{4T} \geqslant \frac{T+1}{8T} \geqslant \frac{1}{8}, \tag{178}$$

having used $\sigma \leqslant (T-1)/2$. Furthermore, notice that by definitions of the instances both $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ will belong to $\mathcal{M}_{2+2\sigma}$. Let $\mathfrak{A}$ be an algorithm, using Lemma C.14, we can lower bound the regret in both the environments $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$:

$$R_{\boldsymbol{\mu}}(\mathfrak{A}, T) \geqslant \sum_{i\neq 1} \overline{\Delta}_{i,\boldsymbol{\mu}}(T,T)\mathbb{E}_{\boldsymbol{\mu}}[N_{i,T}] \geqslant \frac{5}{32}\sum_{i\neq 1}\mathbb{E}_{\boldsymbol{\mu}}[N_{i,T}],$$

where $R_{\boldsymbol{\mu}}(\mathfrak{A}, T)$ is the expected cumulative regret at the time horizon T in the environment $\boldsymbol{\mu}$ and $\mathbb{E}_{\boldsymbol{\mu}}[N_{i,T}]$ is the expected value of the total number of pulls of the suboptimal arm i at the time horizon T, for a fixed policy $\mathfrak{A}$. Similarly for the environment $\boldsymbol{\mu}'$:

$$R_{\boldsymbol{\mu}'}(\mathfrak{A}, T) \geqslant \sum_{i\neq i^*} \overline{\Delta}_{i,\boldsymbol{\mu}'}(T,T)\mathbb{E}_{\boldsymbol{\mu}'}[N_{i,T}] \geqslant \frac{1}{8}\sum_{i\neq i^*}\mathbb{E}_{\boldsymbol{\mu}'}[N_{i,T}],$$

where $R_{\boldsymbol{\mu}'}(\mathfrak{A}, T)$ is the expected cumulative regret at the time horizon T in the environment $\boldsymbol{\mu}'$ and $\mathbb{E}_{\boldsymbol{\mu}'}[N_{i,T}]$ is the expected value of the total number of pulls of the suboptimal arm i at the time horizon T, for a fixed policy $\mathfrak{A}$. For ease of notation we define $N_{i\neq 1,\tau} = \sum_{i\neq 1} N_{i,\tau}$ and $N_{i\neq i*,\tau} = \sum_{i\neq i*} N_{i,\tau}$, for all possible stopping times $\tau \in [\![T]\!]$. Notice that the following holds:

$$\sup_{\boldsymbol{\mu}''\in\mathcal{M}_{2\sigma+2}} R_{\boldsymbol{\mu}''}(\mathfrak{A}, T) \geqslant \max\{R_{\boldsymbol{\mu}}(\mathfrak{A}, T), R_{\boldsymbol{\mu}'}(\mathfrak{A}, T)\} \tag{179}$$

$$\geqslant \frac{R_{\boldsymbol{\mu}}(\mathfrak{A}, T) + R_{\boldsymbol{\mu}'}(\mathfrak{A}, T)}{2} \tag{180}$$

$$\geqslant \frac{\min\{\overline{\Delta}_{i,\boldsymbol{\mu}}(T,T), \overline{\Delta}_{1,\boldsymbol{\mu}'}(T,T)\}}{2}(\mathbb{E}_{\boldsymbol{\mu}}[N_{i\neq 1,T}] + \mathbb{E}_{\boldsymbol{\mu}'}[N_{i\neq i*,T}]) \tag{181}$$

$$\geqslant \frac{\min\{\overline{\Delta}_{i,\boldsymbol{\mu}}(T,T), \overline{\Delta}_{1,\boldsymbol{\mu}'}(T,T)\}}{2}(\mathbb{E}_{\boldsymbol{\mu}}[N_{i\neq 1,\tau_{i*,\sigma}}] + \mathbb{E}_{\boldsymbol{\mu}'}[N_{i\neq i*,\tau_{i*,\sigma}}]) \tag{182}$$

Now we select as a stopping time for the adapted filtration $\tau_{i*,\sigma}$, where $\tau_{i*,\sigma}$ is defined as above, so we can write $\mathbb{E}_{\boldsymbol{\mu}}[N_{i\neq i*,\tau_{i*,\sigma}}] = \mathbb{E}_{\boldsymbol{\mu}'}[N_{i\neq i*,\tau_{i*,\sigma}}]$, since by definition $\tau_{i*,\sigma}$ is the r.v. denoting the time at which the i^*-th arm gets played for the $\frac{\sigma}{2}$-th, i.e., up to which the two instances are indistinguishable. This yield to:

$$\mathbb{E}_{\boldsymbol{\mu}}[N_{i\neq 1,\tau_{i*,\sigma}}] + \mathbb{E}_{\boldsymbol{\mu}'}[N_{i\neq i*,\tau_{i*,\sigma}}] = \mathbb{E}_{\boldsymbol{\mu}}[N_{i\neq 1,\tau_{i*,\sigma}}] + \mathbb{E}_{\boldsymbol{\mu}}[N_{i\neq i*,\tau_{i*,\sigma}}] \tag{183}$$

$$= \mathbb{E}_{\boldsymbol{\mu}}[N_{i\neq 1,\tau_{i*,\sigma}} + N_{i\neq i*,\tau_{i*,\sigma}}] \tag{184}$$

$$\geqslant \mathbb{E}_{\boldsymbol{\mu}}[\tau_{i*,\sigma}], \tag{185}$$

where for the last inequality we used the fact that $N_{i\neq 1,\tau_{i*,\sigma}} + N_{i\neq i*,\tau_{i*,\sigma}} = \sum_{t=1}^{\tau_{i*,\sigma}} \mathbb{1}\{I_t \neq 1\} + \mathbb{1}\{I_t \neq i^*\} \geqslant \sum_{t=1}^{\tau_{i*,\sigma}} 1$. Now, in order to find the lower bound we shall find the bound for $\mathbb{E}_{\boldsymbol{\mu}}[\tau_{i*,\sigma}]$. We recall that by definition $i^* \in \arg\max_{i\in[\![K]\!]\setminus\{1\}} \mathbb{E}_{\boldsymbol{\mu}}[\tau_{i,\sigma}]$, so the following holds:

$$(K-1)\mathbb{E}_{\boldsymbol{\mu}}[\tau_{i*,\sigma}] \geqslant \sum_{i=2}^{K} \mathbb{E}_{\boldsymbol{\mu}}[\tau_{i,\sigma}] = \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{i=2}^{K} \tau_{i,\sigma}\right]. \tag{186}$$

Let us sort the arms in ascending order based on the round in which they get to $\frac{\sigma}{2}$ pulls, i.e., according to the value of $\tau_{\cdot,\sigma}$. In particular, we use the notation:

$$\tau_{[1],\sigma} < \tau_{[2],\sigma} < \cdots < \tau_{[K],\sigma}. \tag{187}$$

Thus, we can rewrite:

$$\sum_{i=2}^{K} \tau_{i,\sigma} \geqslant \sum_{j=1}^{K-1} \tau_{[j],\sigma}, \tag{188}$$

by a pigeonhole-like argument it is easy to infer that $\tau_{[j],\sigma} \geqslant j\frac{\sigma}{2}$, then:

$$\sum_{i=2}^{K} \tau_{i,\sigma} \geqslant \sum_{j=1}^{K-1} \tau_{[j],\sigma} \geqslant \frac{\sigma}{2}\sum_{j=1}^{K-1} j \geqslant \frac{\sigma(K-1)K}{4}. \tag{189}$$

Substituting in (186) we obtain:

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_{i*,\sigma}] \geqslant \frac{K\sigma}{4} \tag{190}$$

Putting all together, from Equation (182):

$$\sup_{\boldsymbol{\mu}''\in\mathcal{M}_{2\sigma+2}} R_{\boldsymbol{\mu}''}(\mathfrak{A}, T) \geqslant \frac{1}{8}\mathbb{E}_{\boldsymbol{\mu}}[\tau_{i*,\sigma}] \geqslant \frac{K}{32}\sigma, \tag{191}$$

the final result follows substituting $\sigma \leftarrow \frac{\sigma-2}{2}$. □

B.4 Proofs of Section 7

Theorem 7.1 (`ET-Beta-SWTS` Bound)**.** *Under Assumption 3.2, for the* `ET-Beta-SWTS` *algorithm with a window of size $\tau \in [\![T]\!]$, for every arm $i \in [\![K]\!]\setminus\{i^*(T)\}$, it holds that:*

$$\mathbb{E}_{\boldsymbol{\mu}}[N_{i,T}] \leqslant O\left(\underbrace{\Gamma}_{(i)} + \underbrace{\frac{T\log(T)}{\tau\Delta_i'^2}}_{(ii)} + \underbrace{(\sigma'-\Gamma)\tau}_{(iii)}\right).$$

Proof. For ease of notation, we set $\sigma'(T;\tau) = \sigma'(\tau)$, $\overline{\mu}_{i*(T)}(\sigma'(T;\tau);\tau) = \overline{\mu}_{i*(T)}(\sigma'(\tau))$ and $\Delta'_i(T;\tau) = \Delta_i$. Using Lemma 5.1 in the form stated in Equation (14), we can bound the expected value of the number of pulls at the time horizon as:

$$\mathbb{E}_\tau[N_{i,T}] \leqslant \Gamma + 1 + l + \frac{\omega T}{\tau} + \underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\left\{p^i_{i,t,\tau} > \frac{1}{T}, I_t = i, N_{i,t,\tau} \geqslant \omega\right\}\right]}_{\text{(A)}}$$

$$+ \underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \left(\frac{1}{p^i_{i*(t),t,\tau}} - 1\right) \mathbb{1}\left\{I_t = i^*(t)\right\}\right]}_{\text{(B)}}.$$

In what follows, we will consider, without loss of generality, the first arm to be the best, i.e., for each round $t \in [\![T]\!]$ we have that $i^*(t) = i^*(T) = 1$. Furthermore we will set $y_{i,t} = y_i = \overline{\mu}_1(\sigma'(\tau)) - \frac{\Delta_i}{3}$, $x_i = \mu_i(T) + \frac{\Delta_i}{3}$, $\omega = \frac{\log T}{(y_i - x_i)^2}$ and $l = (\sigma'(\tau) - \Gamma)(\tau + K)$. We define $\hat{\mu}_{i,t,\tau} = \frac{S_{i,t,\tau}}{N_{i,t,\tau}}$ We further decompose A in two contributions, formally:

$$\text{(A)} = \underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\left\{p^i_{i,t,\tau} > \frac{1}{T}, \hat{\mu}_{i,t,\tau} \leqslant x_i, I_t = i, N_{i,t,\tau} \geqslant \omega\right\}\right]}_{\text{(A1)}}$$

$$+ \underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\left\{p^i_{i,t,\tau} > \frac{1}{T}, \hat{\mu}_{i,t,\tau} > x_i, I_t = i, N_{i,t,\tau} \geqslant \omega\right\}\right]}_{\text{(A2)}}.$$

Term (A2) Focusing on (A2):

$$\text{(A2)} \leqslant \mathbb{E}\left[\sum_{t=K\Gamma+1}^{T} \mathbb{1}\left\{\hat{\mu}_{i,t,T} > x_i, N_{i,t,\tau} \geqslant \omega\right\}\right] \tag{192}$$

$$\leqslant \sum_{t=K\Gamma+1}^{T} \mathbb{P}(\hat{\mu}_{i,t,\tau} > x_i | N_{i,t,\tau} \geqslant \omega) \tag{193}$$

$$\leqslant \sum_{t=K\Gamma+1}^{T} \mathbb{P}(\hat{\mu}_{i,t,\tau} - \mathbb{E}[\hat{\mu}_{i,t,\tau}] > x_i - \mathbb{E}[\hat{\mu}_{i,t,\tau}] | N_{i,t,\tau} \geqslant \omega) \tag{194}$$

$$\leqslant \sum_{t=K\Gamma+1}^{T} \sum_{j=\omega}^{\tau} \mathbb{P}(\hat{\mu}_{i,t,\tau} - \mathbb{E}[\hat{\mu}_{i,t,\tau}] > x_i - \mathbb{E}[\hat{\mu}_{i,t,\tau}],\ N_{i,t,\tau} = j) \tag{195}$$

$$\leqslant \sum_{t=K\Gamma+1}^{T} \sum_{j=\omega}^{\tau} \mathbb{P}(\hat{\mu}_{i,t,\tau} - \mathbb{E}[\hat{\mu}_{i,t,\tau}] > x_i - \mu_i(T),\ N_{i,t,\tau} = j) \tag{196}$$

$$\leqslant \sum_{t=1}^{T} \tau \exp\left(-2\omega(x_i - \mu_i(T))^2\right), \tag{197}$$

where the last inequality follows from the Chernoff-Hoeffding bound (Lemma C.1). By definition it holds that $x_i - \mu_i(T) = y_i - x_i$, substituting we obtain:

$$\sum_{t=1}^{T} \tau \exp\left(-2\omega(x_i - \mu_i(T))^2\right) = \sum_{t=1}^{T} \tau \exp\left(-2\log(T)\right) \leqslant 1. \tag{198}$$

Term (A1) Evaluating the term (A1) we can rewrite:

$$\text{(A1)} \leqslant \mathbb{E}\left[\sum_{t=K\Gamma+1}^{T} \mathbb{1}\left\{\underbrace{p^i_{i,t,\tau} > \frac{1}{T}, N_{i,t,\tau} \geqslant \omega, \hat{\mu}_{i,t,\tau} \leqslant x_i}_{\text{(C)}}\right\}\right].$$

We wish to evaluate if condition (C) ever occurs. To this end, let $(\mathcal{F}_{t-1})_{t\in[\![T]\!]}$ be the canonical filtration. We have:

$$\begin{aligned}
&\mathbb{P}(\theta_{i,t,\tau} > y_i | N_{i,t,\tau} \geqslant \omega, \hat{\mu}_{i,t,\tau} \leqslant x_i, \mathcal{F}_{t-1}) = \\
&= \mathbb{P}\left(\text{Beta}\left(\hat{\mu}_{i,t,\tau} N_{i,t,\tau}+1, (1-\hat{\mu}_{i,t,\tau})N_{i,t,\tau}+1\right) > y_i | \hat{\mu}_{i,t,\tau} \leqslant x_i, N_{i,t,\tau} \geqslant \omega\right) && (199)\\
&\leqslant \mathbb{P}\left(\text{Beta}\left(x_i N_{i,t,\tau}+1, (1-x_i)N_{i,t,\tau}+1\right) > y_i | N_{i,t,\tau} \geqslant \omega\right) && (200)\\
&\leqslant F^B_{N_{i,t,\tau}+1, y_i}\left(x_i N_{i,t,\tau} | N_{i,t,\tau} \geqslant \omega\right) && (201)\\
&\leqslant F^B_{N_{i,t,\tau}, y_i}\left(x_i N_{i,t,\tau} | N_{i,t,\tau} \geqslant \omega\right) && (202)\\
&\leqslant \exp\left(-N_{i,t,\tau} d(x_i, y_i) | N_{i,t,\tau} \geqslant \omega\right) && (203)\\
&\leqslant \exp\left(-2\omega(y_i - x_i)^2\right), && (204)
\end{aligned}$$

where the last-but-one inequality follows from the generalized Chernoff-Hoeffding bounds (Lemma C.1). Equation (201) was derived and the Beta-Binomial identity (Fact C.5), (200) by exploiting that a sample from $\text{Beta}\left(x_i N_{i,t,\tau}+1, (1-x_i)N_{i,t,\tau}+1\right)$ is likely to be as large as a sample from $\text{Beta}(\hat{\mu}_{i,t} N_{i,t,\tau}+1, (1-\hat{\mu}_{i,t,\tau})N_{i,t,\tau}+1)$ (reported formally in Lemma C.17) and finally (202) follows from Lemma C.16. Therefore, for $\omega = \frac{\log(T)}{(y_i - x_i)^2}$, we have:

$$\mathbb{P}(\theta_{i,t,\tau} > | y_i, N_{i,t,\tau} \geqslant \omega, \hat{\mu}_{i,t,\tau} \leqslant x_i, \mathcal{F}_{t-1}) \leqslant \frac{1}{T}, \tag{205}$$

in contradiction with condition (C), so (A1)$= 0$.

Term (B) We now tackle term (B):

$$(\text{B}) = \mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \left(\frac{1}{p^i_{1,t,\tau}} - 1\right) \mathbb{1}\left\{I_t = 1\right\}\right], \tag{206}$$

we notice that from the algorithms design we can infer that for any round $t \geqslant K\Gamma + (\sigma'(\tau) - \Gamma) + 1$ every arm has been played at least $\sigma'(\tau)$ times, since it means that for every interval τ the best arm is played at least once. To take into account this information, we rewrite (B) as follows (so that we do not have to specify that we restrict just to those filtrations such that $N_{1,t} \geqslant \sigma'(\tau)$):

$$\mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \left(\frac{1}{p^i_{1,t,\tau}} - 1\right) \mathbb{1}\left\{I_t = 1\right\}\right] = \mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \left(\frac{1}{p^i_{1,t,\tau}} - 1\right) \mathbb{1}\left\{I_t = 1, N_{i,t} \geqslant \sigma'(\tau)\right\}\right]. \tag{207}$$

Using the properties of the expected value, we write (B) as:

$$(\text{B}) = \mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \left(\frac{1}{p^i_{1,t,\tau}} - 1\right) \mathbb{1}\left\{\underbrace{I_t = 1,\ N_{1,t} \geqslant \sigma'(\tau)}_{\mathcal{C}1}\right\}\right] \tag{208}$$

$$(\text{B}) = \mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\{\mathcal{C}1\} \mathbb{E}\left[\left(\frac{1}{p^i_{1,t,\tau}} - 1\right) \middle| \mathbb{1}\{\mathcal{C}1\}\right]\right], \tag{209}$$

since by definition $N_{1,t,\tau} \leqslant \tau$ we can use a peeling-like argument similar to that in Fiandri et al. (2025) we can decompose condition $\mathcal{C}1$ in $\log(\tau)$ sub-events $\mathcal{C}1_j$ with $j \in [\![1, \log(\tau)]\!]$ defined as:

$$\{\mathcal{C}1_j\} := \{I_t = 1,\ N_{1,t} \geqslant \sigma'(\tau),\ \underbrace{e^{j-1}}_{:=N_{j-1}} < N_{1,t,\tau} \leqslant \underbrace{e^j}_{:=N_j}\}, \tag{210}$$

with:

$$\{\mathcal{C}1_1\} := \{I_t = 1,\ N_{1,t} \geqslant \sigma'(\tau),\ \underbrace{0}_{N_0} \leqslant N_{1,t,\tau} \leqslant \underbrace{e}_{N_1}\}. \tag{211}$$

Thus, we can rewrite (B) as:

$$(\text{B}) = \mathbb{E}\left[\underbrace{\sum_{j=1}^{\log(\tau)} \sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\{\mathcal{C}1_j\} \mathbb{E}\left[\left(\frac{1}{p^i_{1,t,\tau}} - 1\right) \middle| \mathbb{1}\{\mathcal{C}1\}\right]}_{(\text{B}')}\right]. \tag{212}$$

Noticing that for every j the inner summands that will contribute to the total summation are those for which condition $\mathcal{C}1_j$ holds true, we can write:

$$\sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\{\mathcal{C}1_j\}\mathbb{E}\left[\left(\frac{1}{p^i_{1,t,\tau}}-1\right)\middle|\mathbb{1}\{\mathcal{C}1\}\right] = \sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\{\mathcal{C}1_j\}\underbrace{\mathbb{E}\left[\left(\frac{1}{p^i_{1,t,\tau}}-1\right)\middle|\ \mathcal{C}1_j\right]}_{(*)}, \tag{213}$$

now we wish to evaluate $(*)$. We write:

$$(*) = \mathbb{E}_{N_{1,t,\tau},\ \underline{\mu}_1(N_{i,t,\tau})}\left[\underbrace{\mathbb{E}\left[\left(\frac{1}{p^i_{1,t,\tau}}-1\right)\middle|\ \mathcal{C}1_j,\ N_{1,t,\tau},\ \underline{\mu}_1(N_{1,t,\tau})\right]}_{(**)}\right], \tag{214}$$

where the outer expected value runs on every possible instatiation of $N_{1,t,\tau}$ that respect condition $\mathcal{C}1_j$ and every possible set of $N_{1,t,\tau}$ expected rewards of the plays of the best arm at time t. Now, we can bound $(**)$ using Lemma 5.2 so that is possible to exploit the bounds provided for the stationary case in Lemma 2.9 by Agrawal and Goyal (2017). In particular:

$$(**) = \sum_{s=0}^{N_{1,t,\tau}} \frac{f_{N_{1,t,\tau},\ \underline{\mu}_1(N_{1,t,\tau})}(s)}{F^B_{N_{1,t,\tau}+1,y_i}(s)} - 1 \leqslant \sum_{s=0}^{N_{1,t,\tau}} \frac{f_{N_{1,t,\tau},\ \overline{\mu}_1(N_{1,t,\tau})}(s)}{F^B_{N_{1,t,\tau}+1,y_i}(s)} - 1 \tag{215}$$

$$(**) \leqslant \begin{cases} \frac{3}{\Delta_i'} & if\ j \leqslant \log(\frac{8}{\Delta_i'}) \\ \\ \Theta\left(e^{-\frac{\Delta_i'^2 N_{1,t,\tau}}{2}} + \frac{1}{(N_{1,t,\tau}+1)\Delta_i'^2}e^{-D_i N_{1,t,\tau}} + \frac{1}{e^{\frac{\Delta_i'^2 N_{1,t,\tau}}{4}}-1}\right) & if\ j > \log(\frac{8}{\Delta_i'}) \end{cases}, \tag{216}$$

where $\Delta_i' = \overline{\mu}_1(N_{1,t,j}) - y_i \geqslant \overline{\mu}_1(\sigma'(T;\tau);\tau) - y_i$ and $D_i = y_i \log(\frac{y_i}{\overline{\mu}_1(N_{1,t,j})}) + (1-y_i)\log(\frac{1-y_i}{1-\overline{\mu}_1(N_{1,t,j})}) \geqslant 2(\overline{\mu}_1(N_{1,t,j}) - y_i)^2 \geqslant 2(\overline{\mu}_1(\sigma'(T;\tau);\tau) - y_i)^2$, for every round $t \geqslant K\Gamma + l + 1$, where the last inequality follow from Pinsker's inequality. Using the fact that by definition $N_{j-1} \leqslant N_{1,t,\tau} \leqslant N_j$, we can write:

$$(**) \leqslant \begin{cases} \frac{9}{\Delta_i} & if\ j \leqslant \log(\frac{8}{\Delta_i'}) \\ \\ \Theta\left(e^{-\frac{\Delta_i^2 N_{j-1}}{18}} + \frac{9}{(N_{j-1}+1)\Delta_i^2}e^{-\frac{2}{9}\Delta_i^2 N_{j-1}} + \frac{1}{e^{\frac{\Delta_i^2 N_{j-1}}{36}}-1}\right) & if\ j > \log(\frac{8}{\Delta_i'}) \end{cases}, \tag{217}$$

$$\leqslant \begin{cases} \frac{9}{\Delta_i} & if\ j \leqslant \log(\frac{8}{\Delta_i'}) \\ \\ \Theta\left(\frac{18}{\Delta_i^2 N_{j-1}} + \frac{9}{N_{j-1}\Delta_i^2} + \frac{36}{\Delta_i^2 N_{j-1}}\right) & if\ j > \log(\frac{8}{\Delta_i'}), \end{cases} \tag{218}$$

where we exploited the fact that $e^{-x} \leqslant \frac{1}{x}$ for $x > 0$ and $1 + x \leqslant e^x$ for every x. Then for fixed j the bound does not depend neither on $N_{1,t,\tau}$ nor on $\overline{\mu}_1(N_{1,t,j})$, leading to:

$$(*) \leqslant \begin{cases} \frac{9}{\Delta_i} & if\ j \leqslant \log(\frac{8}{\Delta_i'}) \\ \\ \Theta\left(\frac{65}{\Delta_i^2 N_{j-1}}\right) & if\ j > \log(\frac{8}{\Delta_i'}) \end{cases}, \tag{219}$$

that yields to:

$$(\mathrm{B}') \leqslant O\left(\frac{9}{\Delta_i}\underbrace{\sum_{j \leqslant \log(\frac{8}{\Delta_i'})}\sum_{t=K\Gamma+l+1}^{T}\mathbb{1}\{\mathcal{C}1_j\}}_{(\mathrm{B1})} + \sum_{j > \log(\frac{8}{\Delta_i'})}^{\log(\tau)}\frac{1}{N_{j-1}\Delta_i^2}\underbrace{\sum_{t=K\Gamma+l+1}^{T}\mathbb{1}\{\mathcal{C}1_j\}}_{(\mathrm{B2})}\right). \tag{220}$$

First we face term (B2):

$$\begin{aligned}
\text{(B2)} = \sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\{\mathcal{C}1_j\} &\leqslant \sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\{I_t = 1, N_{j-1} < N_{1,t,\tau} \leqslant N_j\} && (221)\\
&\leqslant \sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\{I_t = 1, N_{1,t,\tau} \leqslant N_j\} && (222)\\
&\leqslant \frac{N_j T}{\tau}, && (223)
\end{aligned}$$

where in the last inequality we used Lemma C.13. Now we bound (B1):

$$\begin{aligned}
\text{(B1)} \leqslant \sum_{j \leqslant \log(\frac{8}{\Delta_i'})} \sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\{\mathcal{C}1_j\} = \sum_{t=K\Gamma+l+1}^{T} \sum_{j \leqslant \log(\frac{8}{\Delta_i'})} \mathbb{1}\{\mathcal{C}1_j\} && (224)\\
\leqslant \sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\left\{I_t = 1, N_{1,t,\tau} \leqslant \frac{8}{\Delta_i'}\right\} && (225)\\
\leqslant \frac{8T}{\tau\Delta_i'}, && (226)
\end{aligned}$$

where again the last inequality follows from Lemma C.13. Replacing all the terms we can rewrite (B) as:

$$\text{(B)} \leqslant O\left(\frac{72T}{\tau\Delta_i\Delta_i'} + \sum_{j>\log(\frac{8}{\Delta_i'})}^{\log(\tau)} \frac{N_j T}{N_{j-1}\tau\Delta_i^2}\right), \tag{227}$$

observing that by definition $\Delta_i' \geqslant \Delta_i$ and that $\frac{N_j}{N_{j-1}} = e$ we obtain:

$$\text{(B)} \leqslant O\left(\frac{72T}{\tau\Delta_i^2} + \frac{eT\log(\tau)}{\tau\Delta_i^2}\right), \tag{228}$$

summing all the terms yield to the statement. □

Theorem 7.2 (γ-`ET-SWGTS` Bound)**.** *Under Assumption 3.3, setting* $\gamma \leqslant \min\left\{\frac{1}{4\lambda^2}, 1\right\}$*, for the* γ*-`ET-SWGTS` algorithm with a window of size* $\tau \in [\![T]\!]$*, for every arm* $i \in [\![K]\!]\backslash\{i^*(T)\}$*, it holds that:*

$$\mathbb{E}_{\boldsymbol{\mu}}[N_{i,T}] \leqslant O\left(\underbrace{\Gamma}_{(i)} + \underbrace{\frac{T\log(T\Delta_i'^2 + e^6)}{\gamma\tau\Delta_i'^2} + \frac{T}{\tau}}_{(ii)} + \underbrace{(\sigma' - \Gamma)\tau}_{(iii)}\right).$$

Proof. For ease of notation we set $\sigma'(T;\tau) = \sigma'(\tau), \overline{\mu}_{i^*(T)}(\sigma(T;\tau);\tau) = \overline{\mu}_{i^*(T)}(\sigma'(\tau))$ and $\Delta_i'(T;\tau) = \Delta_i$. Using Lemma B.1, we can bound the expected value of the number pulls of the suboptimal arm at the time horizon as:

$$\begin{aligned}
\mathbb{E}_\tau[N_{i,T}] \leqslant{}& \Gamma + l + \frac{1}{\epsilon_i} + \frac{T}{\tau} + \frac{\omega T}{\tau} + \underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\left\{p_{i,t,\tau}^i > \frac{1}{T\epsilon_i}, I_t = i, N_{i,t,\tau} \geqslant \omega\right\}\right]}_{\text{(A)}}\\
&+ \underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \left(\frac{1}{p_{i^*(t),t,\tau}^i} - 1\right) \mathbb{1}\{I_t = i^*(t)\}\right]}_{\text{(B)}}.
\end{aligned}$$

In what follows, we will consider, without loss of generality, the first arm to be the best, i.e., for each round $t \in [\![T]\!]$ we have that $i^*(t) = i^*(T) = 1$. Furthermore we will set $y_{i,t} = y_i = \overline{\mu}_1(\sigma'(\tau)) - \frac{\Delta_i}{3}$, $x_i = \mu_i(T) + \frac{\Delta_i}{3}$, $\omega = \frac{32\log(T\Delta_i^2 + e^6)}{\gamma(y_i - x_i)^2}$ and $l = (\sigma'(\tau) - \Gamma)(\tau + K)$. We define $\overline{\hat{\mu}}_{i,t,\tau} = \frac{S_{i,t,\tau}}{N_{i,t,\tau}}$ We further decompose (A) in two contributions, formally:

$$(\mathrm{A}) = \underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\left\{p^{i}_{i,t,\tau} > \frac{1}{T\epsilon_i}, \overline{\hat{\mu}}_{i,t,\tau} \leqslant x_i, I_t = i, N_{i,t,\tau} \geqslant \omega\right\}\right]}_{(\mathrm{A1})}$$

$$+ \underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\left\{p^{i}_{i,t,\tau} > \frac{1}{T\epsilon_i}, \overline{\hat{\mu}}_{i,t,\tau} > x_i, I_t = i, N_{i,t,\tau} \geqslant \omega\right\}\right]}_{(\mathrm{A2})}. \tag{229}$$

Let us first tackle term (A2).

Term (A2) Focusing on (A2):

$$(\mathrm{A2}) \leqslant \mathbb{E}\left[\sum_{t=K\Gamma+1}^{T} \mathbb{1}\left\{\overline{\hat{\mu}}_{i,t,T} > x_i, N_{i,t,\tau} \geqslant \omega\right\}\right] \tag{230}$$

$$\leqslant \sum_{t=K\Gamma+1}^{T} \mathbb{P}(\overline{\hat{\mu}}_{i,t,\tau} > x_i | N_{i,t,\tau} \geqslant \omega) \tag{231}$$

$$\leqslant \sum_{K\Gamma+1}^{T} \mathbb{P}(\overline{\hat{\mu}}_{i,t,\tau} - \mathbb{E}[\overline{\hat{\mu}}_{i,t,\tau}] > x_i - \mathbb{E}[\overline{\hat{\mu}}_{i,t,\tau}] | N_{i,t,\tau} \geqslant \omega) \tag{232}$$

$$\leqslant \sum_{t=K\Gamma+1}^{T} \sum_{j=\omega}^{\tau} \mathbb{P}(\overline{\hat{\mu}}_{i,t,\tau} - \mathbb{E}[\overline{\hat{\mu}}_{i,t,\tau}] > x_i - \mathbb{E}[\overline{\hat{\mu}}_{i,t,\tau}], N_{i,t,\tau} = j) \tag{233}$$

$$\leqslant \sum_{t=K\Gamma+1}^{T} \sum_{j=\omega}^{\tau} \mathbb{P}(\overline{\hat{\mu}}_{i,t,\tau} - \mathbb{E}[\overline{\hat{\mu}}_{i,t,\tau}] > x_i - \mu_i(T),\ N_{i,t,\tau} = j) \tag{234}$$

$$\leqslant \sum_{t=1}^{T} \tau \exp\left(-\frac{1}{2\lambda^2}\omega(x_i - \mu_i(T))^2\right), \tag{235}$$

where the last inequality follows from the Chernoff-Hoeffding bound (Lemma C.9). Substituting ω and noticing that $x_i - \mu_i(T) = y_i - x_i$, we obtain:

$$\sum_{t=K\Gamma+1}^{T} \tau \exp\left(-\frac{1}{2\lambda^2}\omega(x_i - \mu_i(T))^2\right) \leqslant \frac{1}{\Delta_i^2}. \tag{236}$$

Term (A1) We focus now on (A1):

$$(\mathrm{A1}) \leqslant \mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\left\{\underbrace{p^{i}_{i,t,\tau} > \frac{1}{T\epsilon_i}, N_{i,t,\tau} \geqslant \omega, \overline{\hat{\mu}}_{i,t,\tau} \leqslant x_i}_{(\mathrm{C})}\right\}\right]. \tag{237}$$

We wish to evaluate if ever condition (C) occurs. In this setting, $\theta_{i,t,\tau}$ is a Gaussian random variable distributed as $\mathcal{N}\left(\overline{\hat{\mu}}_{i,t,\tau}, \frac{1}{\gamma N_{i,t,\tau}}\right)$. We recall that an $\mathcal{N}\left(m, \sigma^2\right)$ distributed r.v. (i.e., a Gaussian random variable with mean m and variance σ^2) is stochastically dominated by $\mathcal{N}\left(m', \sigma^2\right)$ distributed r.v. if $m' \geqslant m$ (Lemma C.17). Therefore, given $\overline{\hat{\mu}}_{i,t,\tau} \leqslant x_i$, the distribution of $\theta_{i,t,\tau}$ is stochastically dominated by $\mathcal{N}\left(x_i, \frac{1}{\gamma N_{i,t,\tau}}\right)$. Formally:

$$\mathbb{P}\left(\theta_{i,t,\tau} > y_i \mid N_{i,t,\tau} > \omega, \overline{\hat{\mu}}_{i,t,\tau} \leqslant x_i, \mathcal{F}_{t-1}\right) \leqslant \mathbb{P}\left(\mathcal{N}\left(x_i, \frac{1}{\gamma N_{i,t,\tau}}\right) > y_i \middle| \mathcal{F}_{t-1}, N_{i,t,\tau} > \omega\right). \tag{238}$$

Using Lemma C.8 we have:

$$\mathbb{P}\left(\mathcal{N}\left(x_i, \frac{1}{\gamma N_{i,t,\tau}}\right) > y_i\right) \leqslant \frac{1}{2} e^{-\frac{(\gamma N_{i,t,\tau})(y_i - x_i)^2}{2}} \tag{239}$$

$$\leqslant \frac{1}{2} e^{-\frac{(\gamma\omega)(y_i - x_i)^2}{2}} \tag{240}$$

which is smaller than $\frac{1}{T\Delta_i^2}$ because $\omega \geqslant \frac{2\log\left(T\Delta_i^2\right)}{\gamma(y_i - x_i)^2}$. Substituting into Equation (238), we get:

$$\mathbb{P}\left(\theta_{i,t,\tau} > y_i \mid N_{i,t,\tau} \geqslant \omega, \overline{\hat{\mu}}_{i,t,\tau} \leqslant x_i, \mathcal{F}_{t-1}\right) \leqslant \frac{1}{T\Delta_i^2}, \tag{241}$$

in contradiction with the condition (C).

Term (B) We notice that by design of the algorithms that every $\tau + K$ rounds, after the forced exploration, every arm will be played at least once, therefore after $K\Gamma + (\sigma'(\tau) - \Gamma)(\tau + K)$ rounds every arm has been surely played at least $\sigma'(\tau)$ times. In order to encapsulate this information without the need to say that we restrict to filtrations for which the aformentioned condition holds, we rewrite (B) in an equivalent form:

$$(\mathrm{B}) = \mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \left(\frac{1 - p^i_{1,t,\tau}}{p^i_{1,t,\tau}}\right) \mathbb{1}\{I_t = 1, N_{1,t} \geqslant \sigma'(\tau)\}\right], \tag{242}$$

we will consider two different contribution to (B):

$$(\mathrm{B}) = \underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \left(\frac{1 - p^i_{1,t,\tau}}{p^i_{1,t,\tau}}\right) \mathbb{1}\left\{\underbrace{I_t = 1, N_{1,t} \geqslant \sigma'(\tau),\ N_{1,t,\tau} \leqslant \omega}_{\mathcal{C}1}\right\}\right]}_{(\mathrm{B1})}$$

$$+ \underbrace{\mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \left(\frac{1 - p^i_{1,t,\tau}}{p^i_{1,t,\tau}}\right) \mathbb{1}\left\{\underbrace{I_t = 1, N_{1,t} \geqslant \sigma'(\tau), N_{1,t,\tau} > \omega}_{\mathcal{C}2}\right\}\right]}_{(\mathrm{B2})}. \tag{243}$$

First we face (B1), doing what we have done in the proof of Theorem 7.1 we rewrite the term:

$$(\mathrm{B1}) = \mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\{\mathcal{C}1\} \mathbb{E}\left[\left(\frac{1}{p^i_{1,t,\tau}} - 1\right) \middle| \mathbb{1}\{\mathcal{C}1\}\right]\right] \tag{244}$$

$$= \mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\{\mathcal{C}1\} \underbrace{\mathbb{E}\left[\left(\frac{1}{p^i_{1,t,\tau}} - 1\right) \middle| \mathcal{C}1\right]}_{(*)}\right], \tag{245}$$

where the last inequality follows from the fact that the only summands that will contribute to the total summation within the outer expected value are those for which $\mathcal{C}1$ holds true. We now wish to evaluate $(*)$:

$$(*) = \mathbb{E}_{N_{i,t,\tau}}\left[\underbrace{\mathbb{E}\left[\left(\frac{1}{p^i_{1,t,\tau}} - 1\right) \middle| \mathcal{C}1,\ N_{1,t,\tau} = j\right]}_{(**)}\right], \tag{246}$$

where the outer expected value runs on every possible number of plays of the optimal arm within a window τ that satisfy condition $\mathcal{C}1$. Given $\mathcal{F}_{t-1}$, let Θ_j denote a $\mathcal{N}\left(\overline{\hat{\mu}}_{1,t,\tau}, \frac{1}{\gamma j}\right)$ distributed Gaussian random variable. Let G_j be the geometric random variable denoting the number of consecutive independent trials until and including the trial where a sample of Θ_j becomes greater than y_i. Then observe that $p^i_{1,t,\tau} = \Pr\left(\Theta_j > y_i \mid \mathcal{F}_{t-1}\right)$ and:

$$\mathbb{E}\left[\frac{1}{p^i_{1,t,\tau}} \middle| \mathcal{C}1,\ N_{1,t,\tau} = j\right] = \mathbb{E}\left[\mathbb{E}\left[G_j \mid \mathcal{F}_{t-1}\right]\right] = \mathbb{E}\left[G_j\right]. \tag{247}$$

Let $z = \sqrt{\log r}$ and let random variable MAX $_r$ denote the maximum of r independent samples of Θ_j. We abbreviate $\overline{\hat{\mu}}_{1,t,\tau}$ to $\overline{\hat{\mu}}_1$ and we will abbreviate $\overline{\mu}_1(\sigma'(\tau))$ as μ_1 and $\overline{\Delta}'_i(T, \tau)$ as Δ_i in the following. Then for any integer $r \geqslant 1$:

$$\begin{aligned}
\mathbb{P}\left(G_j \leqslant r\right) &\geqslant \mathbb{P}\left(\mathrm{MAX}_r > y_i\right) && (248)\\
&\geqslant \mathbb{P}\left(\mathrm{MAX}_r > \widehat{\overline{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant y_i\right) && (249)\\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(\mathrm{MAX}_r > \widehat{\overline{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant y_i\right) \middle| \mathcal{F}_{t-1}\right]\right] && (250)\\
&= \mathbb{E}\left[\mathbb{1}\left(\widehat{\overline{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant y_i\right) \mathbb{P}\left(\mathrm{MAX}_r > \widehat{\overline{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \middle| \mathcal{F}_{t-1}\right)\right]. && (251)
\end{aligned}$$

For any instantiation F_{t-1} of $\mathcal{F}_{t-1}$, since Θ_j is Gaussian $\mathcal{N}\left(\widehat{\overline{\mu}}_1, \frac{1}{\gamma j}\right)$ distributed r.v., this gives using C.7:

$$\begin{aligned}
\mathbb{P}\left(\mathrm{MAX}_r > \widehat{\overline{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \middle| \mathcal{F}_{t-1} = F_{t-1}\right) &\geqslant 1 - \left(1 - \frac{1}{\sqrt{2\pi}} \frac{z}{(z^2+1)} e^{-z^2/2}\right)^r && (252)\\
&= 1 - \left(1 - \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\log r}}{(\log r + 1)} \frac{1}{\sqrt{r}}\right)^r && (253)\\
&\geqslant 1 - e^{-\frac{r}{\sqrt{4\pi r \log r}}}. && (254)
\end{aligned}$$

For $r \geqslant e^{12}$:

$$\mathbb{P}\left(\mathrm{MAX}_r > \widehat{\overline{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \middle| \mathcal{F}_{t-1} = F_{t-1}\right) \geqslant 1 - \frac{1}{r^2}. \tag{255}$$

Substituting we obtain:

$$\begin{aligned}
\mathbb{P}\left(G_j \leqslant r\right) &\geqslant \mathbb{E}\left[\mathbb{1}\left(\widehat{\overline{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant y_i\right)\left(1 - \frac{1}{r^2}\right)\right] && (256)\\
&= \left(1 - \frac{1}{r^2}\right) \mathbb{P}\left(\widehat{\overline{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant y_i\right). && (257)
\end{aligned}$$

Applying Lemma C.9 to the second term, we can write, since $\mathbb{E}[\widehat{\overline{\mu}}_1] \geqslant \mu_1$ for $t \geqslant K\Gamma + l + 1$:

$$\mathbb{P}\left(\widehat{\overline{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant \mu_1\right) \geqslant 1 - e^{-\frac{z^2}{2\gamma\lambda^2}} \geqslant 1 - \frac{1}{r^2}, \tag{258}$$

being $\gamma \leqslant \frac{1}{4\lambda^2}$. Using, $y_i \leqslant \mu_1$, this gives

$$\mathbb{P}\left(\widehat{\overline{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} \geqslant y_i\right) \geqslant 1 - \frac{1}{r^2}. \tag{259}$$

Substituting all back we obtain:

$$\begin{aligned}
\mathbb{E}\left[G_j\right] &= \sum_{r=0}^{+\infty} \mathbb{P}\left(G_j \geqslant r\right) && (260)\\
&= 1 + \sum_{r=1}^{\infty} \mathbb{P}\left(G_j \geqslant r\right) && (261)\\
&\leqslant 1 + e^{12} + \sum_{r \geqslant 1}\left(\frac{1}{r^2} + \frac{1}{r^2}\right) && (262)\\
&\leqslant 1 + e^{12} + 2 + 2. && (263)
\end{aligned}$$

This shows a constant bound for $(**) = \mathbb{E}\left[G_j\right] - 1 \leqslant e^{12} + 5$ for all j satisfying condition $\mathcal{C}1$. This yields to:

$$(*) \leqslant e^{12} + 5, \tag{264}$$

then (B1) can be written as:

$$\text{(B1)} \leqslant (e^{12} + 5)\mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\{\mathcal{C}1\}\right], \tag{265}$$

where the inner term can be bounded by Lemma C.13, thus obtaining:

$$\text{(B1)} \leqslant (e^{12} + 5)\frac{\omega T}{\tau}. \tag{266}$$

We derive a bound for large j. Let us now consider (B2), making the same steps we have done earlier we can rewrite it as:

$$(\mathrm{B2}) = \mathbb{E}\left[\sum_{t=K\Gamma+l+1}^{T} \mathbb{1}\{\mathcal{C}2\} \underbrace{\mathbb{E}\left[\left(\frac{1}{p^i_{1,t,\tau}} - 1\right)\middle|\mathcal{C}2\right]}_{(*)}\right], \tag{267}$$

and similarly:

$$(*) = \mathbb{E}_{N_{i,t,\tau}}\left[\underbrace{\mathbb{E}\left[\left(\frac{1}{p^i_{1,t,\tau}} - 1\right)\middle|\ \mathcal{C}2,\ N_{1,t,\tau} = j\right]}_{(**)}\right], \tag{268}$$

where the outer expected value runs on every possible number of plays of the optimal arm within a window τ, $N_{i,t,\tau} = j$ that satisfy condition $\mathcal{C}2$. Given $\mathcal{F}_{t-1}$, let Θ_j denote a $\mathcal{N}\left(\overline{\hat{\mu}}_{1,t,\tau}, \frac{1}{\gamma j}\right)$ distributed Gaussian random variable. Let G_j be the geometric random variable denoting the number of consecutive independent trials until and including the trial where a sample of Θ_j becomes greater than y_i. Then observe that $p^i_{1,t,\tau} = \Pr\left(\Theta_j > y_i \mid \mathcal{F}_{t-1}\right)$ and:

$$\mathbb{E}\left[\frac{1}{p^i_{1,t,\tau}}\middle|\ \mathcal{C}2,\ N_{1,t,\tau} = j\right] = \mathbb{E}\left[\mathbb{E}\left[G_j \mid \mathcal{F}_{t-1}\right]\right] = \mathbb{E}\left[G_j\right]. \tag{269}$$

Let $z = \sqrt{\log r}$ and let random variable MAX_r denote the maximum of r independent samples of Θ_j. We abbreviate $\overline{\hat{\mu}}_{1,t,\tau}$ to $\overline{\hat{\mu}}_1$ and we will abbreviate $\overline{\mu}_1(\sigma'(\tau))$ as μ_1 and $\overline{\Delta}_i'(T,\tau)$ as Δ_i in the following. Then for any integer $r \geqslant 1$:

$$\begin{aligned}
\mathbb{P}\left(G_j \leqslant r\right) &\geqslant \mathbb{P}\left(\mathrm{MAX}_r > y_i\right) && (270)\\
&\geqslant \mathbb{P}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6} \geqslant y_i\right) && (271)\\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6} \geqslant y_i\right)\middle|\ \mathcal{F}_{t-1}\right]\right] && (272)\\
&= \mathbb{E}\left[\mathbb{1}\left(\overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} + \frac{\Delta_i}{6} \geqslant \mu_1\right)\mathbb{P}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6}\middle|\ \mathcal{F}_{t-1}\right)\right]. && (273)
\end{aligned}$$

where we used that $y_i = \mu_1 - \frac{\Delta_i}{3}$. Now, since $j \geqslant \omega = \frac{288\log\left(T\Delta_i^2 + e^6\right)}{\gamma\Delta_i^2}$,

$$2\frac{\sqrt{2\log\left(T\Delta_i^2 + e^6\right)}}{\sqrt{\gamma j}} \leqslant \frac{\Delta_i}{6}. \tag{274}$$

Therefore, for $r \leqslant \left(T\Delta_i^2 + e^6\right)^2$,

$$\frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6} = \frac{\sqrt{\log(r)}}{\sqrt{\gamma j}} - \frac{\Delta_i}{6} \leqslant -\frac{\Delta_i}{12}. \tag{275}$$

Then, since Θ_j is $\mathcal{N}\left(\overline{\hat{\mu}}_{1,t,\tau}, \frac{1}{\gamma j}\right)$ distributed random variable, using the upper bound in Lemma C.8, we obtain for any instantiation F_{t-1} of history $\mathbb{F}_{t-1}$,

$$\mathbb{P}\left(\Theta_j > \overline{\hat{\mu}}_{1,t,\tau} - \frac{\Delta_i}{12}\middle|\ \mathcal{F}_{t-1} = F_{t-1}\right) \geqslant 1 - \frac{1}{2}e^{-\gamma j \frac{\Delta_i^2}{288}} \geqslant 1 - \frac{1}{2\left(T\Delta_i^2 + e^6\right)}. \tag{276}$$

being $j \geqslant \omega$. This implies:

$$\mathbb{P}\left(\mathrm{MAX}_r > \overline{\hat{\mu}}_{1,t,\tau} + \frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6}\middle|\ \mathcal{F}_{t-1} = F_{t-1}\right) \geqslant 1 - \frac{1}{2^r\left(T\Delta_i^2 + e^6\right)^r}. \tag{277}$$

Also, using Lemma C.9, as $\mathbb{E}[\overline{\widehat{\mu}}_1] \geqslant \mu_1$ for $t \geqslant K\Gamma + l + 1$ we get:

$$\mathbb{P}\left(\overline{\widehat{\mu}}_1 + \frac{z}{\sqrt{\gamma j}} - \frac{\Delta_i}{6} \geqslant y_i\right) \geqslant \mathbb{P}\left(\overline{\widehat{\mu}}_1 \geqslant \mu_1 - \frac{\Delta_i}{6}\right) \geqslant 1 - e^{-j\Delta_i^2/72\lambda^2} \geqslant 1 - \frac{1}{\left(T\Delta_i^2 + e^6\right)^{16}}. \tag{278}$$

Let $T' = \left(T\Delta_i^2 + e^6\right)^2$. Therefore, for $1 \leqslant r \leqslant T'$, we have:

$$\mathbb{P}\left(G_j \leqslant r\right) \geqslant 1 - \frac{1}{2^r \left(T'\right)^{r/2}} - \frac{1}{\left(T'\right)^8}. \tag{279}$$

When $r \geqslant T' \geqslant e^{12}$, we obtain:

$$\mathbb{P}\left(G_j \leqslant r\right) \geqslant 1 - \frac{1}{r^2} - \frac{1}{r^2}. \tag{280}$$

Combining all the bounds we have derived:

$$\begin{align}
\mathbb{E}\left[G_j\right] &\leqslant \sum_{r=0}^{\infty} \mathbb{P}\left(G_j \geqslant r\right) \tag{281}\\
&\leqslant 1 + \sum_{r=1}^{T'} \mathbb{P}\left(G_j \geqslant r\right) + \sum_{r=T'}^{\infty} \mathbb{P}\left(G_j \geqslant r\right) \tag{282}\\
&\leqslant 1 + \sum_{r=1}^{T'} \frac{1}{\left(2\sqrt{T'}\right)^r} + \frac{1}{\left(T'\right)^7} + \sum_{r=T'}^{\infty} \frac{1}{r^2} + \frac{1}{r^{1.5}} \tag{283}\\
&\leqslant 1 + \frac{1}{\sqrt{T'}} + \frac{1}{\left(T'\right)^7} + \frac{2}{T'} + \frac{3}{\sqrt{T'}} \tag{284}\\
&\leqslant 1 + \frac{5}{T\Delta_i^2 + e^6}. \tag{285}
\end{align}$$

This yields to:

$$\text{(B2)} \leqslant O\left(\frac{1}{\Delta_i^2}\right). \tag{286}$$

by summing all the terms we obtain the result from the statement. □

C AUXILIARY LEMMAS

In this section, we report some results that already exist in the bandit literature and have been used to demonstrate our results.

Lemma C.1 (Generalized Chernoff-Hoeffding bound from Agrawal and Goyal (2017))**.** *Let $X_1, \dots, X_n$ be independent Bernoulli random variables with $\mathbb{E}[X_i] = p_i$, consider the random variable $X = \frac{1}{n}\sum_{i=1}^n X_i$, with $\mu = \mathbb{E}[X]$. For any $0 < \lambda < 1 - \mu$ we have:*

$$\mathbb{P}(X \geqslant \mu + \lambda) \leqslant \exp\big(-nd(\mu+\lambda, \mu)\big),$$

and for any $0 < \lambda < \mu$

$$\mathbb{P}(X \leqslant \mu - \lambda) \leqslant \exp\big(-nd(\mu-\lambda, \mu)\big),$$

where $d(a,b) := a \log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}$.

Lemma C.2 (Change of Measure Argument from Lattimore and Szepesvári (2020))**.** *Let $(\Omega, \mathcal{F})$ be a measurable space, and $P, Q : \mathcal{F} \to [0,1]$. Let $a < b$ and $X \to [a,b]$ be a $\mathcal{F}$-measurable random variable, we have:*

$$\left|\int_\Omega X(\omega)\, dP(\omega) \; - \int_\Omega X(\omega)\, dQ(\omega)\right| \leqslant (b-a)\delta_{TV}(P,Q). \tag{287}$$

Lemma C.3 (Lattimore and Szepesvári (2020), proposition 2.8)**.** *For a nonnegative random variable X, the expected value $\mathbb{E}[X]$ can be computed as:*

$$\mathbb{E}[X] = \int_0^{+\infty} \mathbb{P}(X > y)\mathrm{d}y.$$

Lemma C.4 (Roos (2001), Theorem 2)**.** *Let us define $\underline{\mu}_n := (\mu_1, \dots, \mu_n)$, $s \in (0, \dots, n)$ and $\mu \in (0,1)$. We have that the total variation distance between two variables $PB(\underline{\mu}_n)$ and $B_s(n,\mu)$ is:*

$$\delta_{TV}(PB(\underline{\mu}_n), B_s(n,\mu)) \leqslant \begin{cases} C_1(s)\theta(\mu,\underline{\mu}_n)^{\frac{s+1}{2}} \frac{\left(1-\frac{s}{s+1}\sqrt{\theta(\mu,\underline{\mu}_n)}\right)}{(1-\sqrt{\theta(\mu,\underline{\mu}_n)})^2} & \textit{if } \theta(\mu,\underline{\mu}_n) < 1 \\ C_2(s)\eta(\mu,\underline{\mu}_n)^{\frac{s+1}{2}}(1+\sqrt{2\eta(\mu,\underline{\mu}_n)})\exp(2\eta(\mu,\underline{\mu}_n)) & \textit{otherwise} \end{cases}, \tag{288}$$

where $\theta(\mu,\underline{\mu}_n) := \frac{\eta(\mu,\underline{\mu}_n)}{2n\mu(1-\mu)}$, $\eta(\mu,\underline{\mu}_n) := 2\gamma_2(\mu,\underline{\mu}_n) + \gamma_1(\mu,\underline{\mu}_n)^2$, $\gamma_k(\mu,\underline{\mu}_n) := \sum_{n'=1}^n (\mu - \mu_{n'})^k$, $C_1(s) := \frac{\sqrt{e}(s+1)^{\frac{1}{4}}}{2}$, $C_2(s) := \frac{(2\pi)^{\frac{1}{4}}\exp\left(\frac{1}{24(s+1)}\right)2^{\frac{s-1}{2}}}{\sqrt{s!}(s+1)^{\frac{1}{4}}}$. To obtain the binomial distribution is sufficient to set $s = 0$ in $B_s(n,\mu)$.

Lemma C.5 (Beta-Binomial identity)**.** *For all positive integers $\alpha, \beta \in \mathbb{N}$, the following equality holds:*

$$F^{beta}_{\alpha,\beta}(y) = 1 - F^{B}_{\alpha+\beta-1,y}(\alpha - 1), \tag{289}$$

where $F^{beta}_{\alpha,\beta}(y)$ is the cumulative distribution function of a beta with parameters α and β, and $F^{B}_{\alpha+\beta-1,y}(\alpha-1)$ is the cumulative distribution function of a binomial variable with $\alpha+\beta-1$ trials having each probability y.

Lemma C.6 (Boland et al. (2002), Theorem 1 (iii))**.** *Let $Y \sim Bin(n,\lambda)$ and $X = \sum X_i$ where the $X_i \sim Bin(n_i, \lambda_i)$ are independent random variables for $i = 1, \dots, k$ then:*

$$X \geqslant_{st} Y \textit{if and only if } \lambda \leqslant \overline{\lambda}_g, \tag{290}$$

$$X \leqslant_{st} Y \textit{if and only if } \lambda \geqslant \overline{\lambda}_{cg}, \tag{291}$$

where $X \geqslant_{st} Y$ means that X is greater than Y in the stochastic order, i.e. $\mathbb{P}(X > m) \geqslant \mathbb{P}(Y > m)$ $\forall m$, and:

$$\overline{\lambda}_g = \left(\prod_{i=1}^k \lambda_i^{n_i}\right)^{\frac{1}{n}}, \tag{292}$$

$$\overline{\lambda}_{cg} = 1 - \left(\prod_{i=1}^k (1-\lambda_i)^{n_i}\right)^{\frac{1}{n}}. \tag{293}$$

Lemma C.7 (Abramowitz and Stegun (1968) Formula 7.1.13)**.** *Let Z be a Gaussian random variable with mean μ and standard deviation σ, then:*

$$\mathbb{P}(Z > \mu + x\sigma) \geqslant \frac{1}{\sqrt{2\pi}}\frac{x}{x^2+1}e^{-\frac{x^2}{2}} \tag{294}$$

Lemma C.8 (Abramowitz and Stegun (1968)). *Let Z be a Gaussian r.v. with mean m and standard deviation σ, then:*

$$\frac{1}{4\sqrt{\pi}}e^{-7z^2/2} < \mathbb{P}(|Z-m| > z\sigma) \leqslant \frac{1}{2}e^{-z^2/2}. \tag{295}$$

Lemma C.9 (Rigollet and Hütter (2023) Corollary 1.7). *Let $X_1, \dots, X_n$ be n independent random variables such that $X_i \sim$ Subg(σ^2), then for any $a \in \mathbb{R}^n$, we have*

$$\mathbb{P}\left[\sum_{i=1}^{n} a_i X_i > t\right] \leqslant \exp\left(-\frac{t^2}{2\sigma^2|a|_2^2}\right), \tag{296}$$

and

$$\mathbb{P}\left[\sum_{i=1}^{n} a_i X_i < -t\right] \leqslant \exp\left(-\frac{t^2}{2\sigma^2|a|_2^2}\right) \tag{297}$$

Of special interest is the case where $a_i = 1/n$ for all i we get that the average $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, satisfies

$$\mathbb{P}(\bar{X} > t) \leqslant e^{-\frac{nt^2}{2\sigma^2}} \quad \text{and} \quad \mathbb{P}(\bar{X} < -t) \leqslant e^{-\frac{nt^2}{2\sigma^2}}$$

Lemma C.10 (Samuels (1965),Tang and Tang (2023), Theorem 2.1 (2)). *Let $X \sim \mathrm{PB}(p_1, \dots, p_n)$, and $\bar{X} \sim \mathrm{Bin}(n, \bar{p})$ with $\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i$, for any convex function $g : [n] \to \mathbb{R}$ in the sense that $g(k+2) - 2g(k+1) + g(k) > 0$ for all $0 \leqslant k \leqslant n-2$, we have:*

$$\mathbb{E}g(X) \leqslant \mathbb{E}g(\bar{X}), \tag{298}$$

where the equality holds if and only if $p_1 = \dots = p_n$ of the Poisson-binomial distribution are all equal to $\bar{p}$ of the binomial distribution.

Lemma C.11 (Johnson and Goldschmidt (2006) Definition 1.2, Hoggar (1974)). *A random variable V taking values in $\mathbb{Z}_+$ is discrete log-concave if its probability mass function $p_V(i) = P(V = i)$ forms a log-concave sequence. That is, V is log-concave if for every $i \geqslant 1$:*

$$p_V(i)^2 \geqslant p_V(i-1)p_V(i+1) \tag{299}$$

Any Bernoulli random variable (i.e., only taking values in $\{0, 1\}$) is discrete log-concave. Furthermore, any binomial distribution is discrete log-concave. In fact, any random variable $S = \sum_{i=1}^{n} X_i$, where X_i are independent (not necessarily identically distributed) Bernoulli variables, is discrete log-concave. Notice then that, by definition $\frac{1}{p_V(i)}$, is discrete log-convex.

Lemma C.12 (Hill and Houdré (1999), Theorem 2 p.152, Remark 13 p.153, Remark 1 p.150). *Let $1 \leqslant \alpha < r \leqslant +\infty$ and let $q : \mathbb{Z} \to [0, +\infty]$ be r-concave (Definition 1 p.150 Hill and Houdré (1999); we highlight that for Remark 1 p.150 Hill and Houdré (1999) to be ∞-concave is equivalent to be discrete log-concave). Then, $\mathcal{J}^\alpha q$ is $(r-\alpha)$-concave, we assume $r - \alpha = +\infty$ when $r = +\infty$ and $r > \alpha$. Where the α-fractional (tail) sum of a function $q : \mathbb{Z} \to [0, \infty]$ is defined for every $\alpha > 0$ by the formula:*

$$\mathcal{J}^\alpha q(n) = \sum_{k=0}^{\infty} \binom{\alpha+k-1}{k} q(n+k), \tag{300}$$

so that for a binomial pdf p_{bin}, being p_{bin} discrete log-concave (see Lemma C.11), follows that $\mathcal{J}^\alpha p_{bin}(-n)$ is ∞-concave on $\mathbb{Z}$ for $\alpha \geqslant 1$.

Lemma C.13 (Combes and Proutiere (2014), Lemma D.1). *Let $A \subset \mathbb{N}$, and $\tau \in \mathbb{N}$ fixed. Define $a(n) = \sum_{t=n-\tau}^{n-1} \mathbb{1}\{t \in A\}$. Then for all $T \in \mathbb{N}$ and $s \in \mathbb{N}$ we have the inequality:*

$$\sum_{n=1}^{T} \mathbb{1}\{n \in A, a(n) \leqslant s\} \leqslant s\lceil T/\tau \rceil. \tag{301}$$

Lemma C.14 (Fiandri et al. (2024), Lemma 4.1). *Let $T \in \mathbb{N}$ be the time horizon, π a fixed policy, and $\boldsymbol{\mu}$ a rising rested bandit. Then, it holds that:*

$$R_{\boldsymbol{\mu}}(\pi, T) \geqslant \sum_{i \neq i^\star(T)} \overline{\Delta}_i(T, T)\mathbb{E}_{\boldsymbol{\mu}}^{\pi}[N_{i,T}], \tag{302}$$

Lemma C.15 (Bretagnolle-Huber inequality and List of KL-Divergences by Gil et al. (2013))**.** *Let P and Q be two probability distributions on a measurable space $(\Omega, \mathcal{F})$. Recall that the total variation between P and Q is defined by*

$$\delta_{TV}(P, Q) = \sup_{A \in \mathcal{F}} \{|P(A) - Q(A)|\}$$

The Kullback-Leibler divergence is defined as follows:

$$D_{\mathrm{KL}}(P \| Q) = \begin{cases} \int_\Omega \log\left(\frac{dP}{dQ}\right) dP & \text{if } P \ll Q \\ +\infty & \text{otherwise} \end{cases}$$

In the above, the notation $P \ll Q$ stands for absolute continuity of P with respect to Q, and $\frac{dP}{dQ}$ stands for the Radon-Nikodym derivative of P with respect to Q.

The Bretagnolle-Huber inequality states:

$$\delta_{TV}(P, Q) \leqslant \sqrt{1 - \exp\left(-D_{\mathrm{KL}}(P \| Q)\right)} \leqslant 1 - \frac{1}{2} \exp\left(-D_{\mathrm{KL}}(P \| Q)\right).$$

A comprehensive list of KL-Divergences is provided in Figure 8. In Figure 8, $\Gamma(x)$, $\psi(x)$, and $B(x)$ denote the Gamma, Digamma, and the multivariate Beta functions, respectively. Also, $\gamma \approx 0.5772$ is the Euler-Mascheroni constant.

Lemma C.16 (Theorem 4.2.3, Example 4.2.4 Roch (2024))**.** *Let $F_{n,p}$ be the CDF of a Bin(n, p) distributed Random Variable, then holds for $m \leqslant n$ and $q \leqslant p$:*

$$F_{n,p}(x) \leqslant F_{m,q}(x) \tag{303}$$

for all x.

Lemma C.17 (Beta and Normal Ordering)**.**

1. *A $\mathcal{N}\left(m, \sigma^2\right)$ distributed r.v. (i.e., a Gaussian random variable with mean m and variance σ^2) is stochastically dominated by $\mathcal{N}\left(m', \sigma^2\right)$ distributed r.v. if $m' \geqslant m$.*
2. *A Beta(α, β) random variable is stochastically dominated by Beta(α', β') if $\alpha' \geqslant \alpha$ and $\beta' \leqslant \beta$.*

These lemmas are informally stated in Agrawal and Goyal (2017), and though intuitive we had difficulties to find formal proofs, so we provide one here.

Proof. Let us consider the ratio of the pdfs of the normal distributed Random Variables:

$$\frac{f_{m',\sigma}(x)}{f_{m,\sigma}(x)} = \exp\left(\frac{-(x-m')^2 + (x-m)^2}{2\sigma^2}\right) = \exp\left(\frac{2x(m'-m) - m'^2 + m^2}{2\sigma^2}\right), \tag{304}$$

that is increasing in $x \in \mathbb{R}$. Similarly let us consider the ratio of the pdfs of two beta distributed random variables:

$$\frac{f_{\alpha',\beta'}(x)}{f_{\alpha,\beta}(x)} = c_{\alpha,\alpha',\beta,\beta'} x^{(\alpha'-\alpha)}(1-x)^{\beta'-\beta}, \tag{305}$$

that is increasing in $x \in (0, 1)$, as $c_{\alpha,\alpha',\beta,\beta'}$ is a constant independent from x. So for $x_0 \leqslant x_1$ for the ratios hold:

$$\begin{cases} \frac{f_{m',\sigma}(x_0)}{f_{m,\sigma}(x_0)} \leqslant \frac{f_{m',\sigma}(x_1)}{f_{m,\sigma}(x_1)} \\ \\ f_{m',\sigma}(x_0) f_{m,\sigma}(x_1) \leqslant f_{m,\sigma}(x_0) f_{m',\sigma}(x_1) \end{cases}, \tag{306}$$

and

$$\begin{cases} \frac{f_{\alpha',\beta'}(x_0)}{f_{\alpha,\beta}(x_0)} \leqslant \frac{f_{\alpha',\beta'}(x_1)}{f_{\alpha,\beta}(x_1)}. \\ \\ f_{\alpha',\beta'}(x_0) f_{\alpha,\beta}(x_1) \leqslant f_{\alpha,\beta}(x_0) f_{\alpha',\beta'}(x_1) \end{cases}. \tag{307}$$

Let us first tackle the normal distribute random variables. Integrating with respect to x_0 up to x_1 we obtain:

$$f_{m,\sigma}(x_1) \int_{-\infty}^{x_1} f_{m',\sigma}(x_0) dx_0 \leqslant f_{m',\sigma}(x_1) \int_{-\infty}^{x_1} f_{m,\sigma}(x_0) dx_0 \tag{308}$$

Name	$D_{KL}\left(f_i \Vert f_j\right)$
Beta	$\log \dfrac{B\left(a_j, b_j\right)}{B\left(a_i, b_i\right)} + \psi\left(a_i\right)\left(a_i - a_j\right) + \psi\left(b_i\right)\left(b_i - b_j\right)$ $+ \left[a_j + b_j - \left(a_i + b_i\right)\right] \psi\left(a_i + b_i\right)$
Chi	$\frac{1}{2}\psi\left(k_i/2\right)\left(k_i - k_j\right) + \log\left[\left(\frac{\sigma_j}{\sigma_i}\right)^{k_j} \frac{\Gamma(k_j/2)}{\Gamma(k_i/2)}\right] + \frac{k_i}{2\sigma_j^2}\left(\sigma_i^2 - \sigma_j^2\right)$
χ^2	$\log \frac{\Gamma(d_j/2)}{\Gamma(d_i/2)} + \frac{d_i - d_j}{2}\psi\left(d_i/2\right)$
Cramér	$\frac{\theta_i + \theta_j}{\theta_i - \theta_j} \log \frac{\theta_i}{\theta_j} - 2$
Dirichlet	$\log \frac{B(\boldsymbol{a}_j)}{B(\boldsymbol{a}_i)} + \sum_{k=1}^{d}\left[a_{i_k} - a_{j_k}\right]\left[\psi\left(a_{i_k}\right) - \psi\left(\sum_{k=1}^{d} a_{i_k}\right)\right]$
Exponential	$\log \frac{\lambda_i}{\lambda_j} + \frac{\lambda_j - \lambda_i}{\lambda_i}$
Gamma	$\left(\frac{\theta_i - \theta_j}{\theta_j}\right) k_i + \log\left(\frac{\Gamma(k_j)\theta_j^{k_j}}{\Gamma(k_i)\theta_i^{k_i}}\right) + \left(k_i - k_j\right)\left(\log\theta_i + \psi\left(k_i\right)\right)$
Multivariate Gaussian	$\frac{1}{2}\left(\log \frac{\lvert\Sigma_j\rvert}{\lvert\Sigma_i\rvert} + \operatorname{tr}\left(\Sigma_j^{-1}\Sigma_i\right)\right) + \frac{1}{2}\left[\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)' \Sigma_j^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right) - n\right]$
Univariate Gaussian	$\frac{1}{2\sigma_j^2}\left[\left(\mu_i - \mu_j\right)^2 + \sigma_i^2 - \sigma_j^2\right] + \log \frac{\sigma_j}{\sigma_i}$
Special Bivariate Gaussian	$\frac{1}{2}\log\left(\frac{1-\rho_j^2}{1-\rho_i^2}\right) + \frac{\rho_j^2 - \rho_j\rho_i}{1-\rho_j^2}$
General Gumbel	$\log \frac{\beta_j}{\beta_i} + \gamma\left(\frac{\beta_i}{\beta_j} - 1\right) + e^{(\mu_j - \mu_i)/\beta_j}\Gamma\left(\frac{\beta_i}{\beta_j} + 1\right) - 1$
Half-Normal	$\log\left(\frac{\sigma_j}{\sigma_i}\right) + \frac{\sigma_i^2 - \sigma_j^2}{2\sigma_j^2}$
Inverse Gaussian	$\frac{1}{2}\left(\frac{\lambda_j}{\lambda_i} + \log\left(\frac{\lambda_i}{\lambda_j}\right) + \frac{\lambda_j(\mu_i - \mu_j)^2}{\mu_i\mu_j^2} - 1\right)$
Laplace	$\log \frac{\lambda_j}{\lambda_i} + \frac{\lvert\theta_i - \theta_j\rvert}{\lambda_j} + \frac{\lambda_i}{\lambda_j}\exp\left(-\frac{\lvert\theta_i - \theta_j\rvert}{\lambda_i}\right) - 1$
Lévy Equal Supports ($\mu_i = \mu_j$)	$\frac{1}{2}\log\left(\frac{c_i}{c_j}\right) + \frac{c_j - c_i}{2c_i}$
Log-normal	$\frac{1}{2\sigma_j^2}\left[\left(\mu_i - \mu_j\right)^2 + \sigma_i^2 - \sigma_j^2\right] + \log \frac{\sigma_j}{\sigma_i}$
Maxwell Boltzmann	$3\log\left(\frac{\sigma_j}{\sigma_i}\right) + \frac{3\left(\sigma_i^2 - \sigma_j^2\right)}{2\sigma_j^2}$
Pareto	$\log\left(\frac{m_i}{m_j}\right)^{a_j} + \log \frac{a_i}{a_j} + \frac{a_j - a_i}{a_i}$, for $m_i \geqslant m_j$ and ∞ otherwise.
Rayleigh	$2\log\left(\frac{\sigma_j}{\sigma_i}\right) + \frac{\sigma_i^2 - \sigma_j^2}{\sigma_j^2}$
Uniform	$\log \frac{b_j - a_j}{b_i - a_i}$ for $(a_i, b_i) \subseteq (a_j, b_j)$ and ∞ otherwise.
General Weibull	$\log\left(\frac{k_i}{k_j}\left[\frac{\lambda_j}{\lambda_i}\right]^{k_j}\right) + \gamma\frac{k_j - k_i}{k_i} + \left(\frac{\lambda_i}{\lambda_j}\right)^{k_j}\Gamma\left(1 + \frac{k_j}{k_i}\right) - 1$

Figure 8: KL Divergences

$$f_{m,\sigma}(x_1)F_{m',\sigma}(x_1) \leqslant f_{m',\sigma}(x_1)F_{m,\sigma}(x_1), \tag{309}$$

so for any $x \in \mathbb{R}$ it holds:

$$\frac{F_{m',\sigma}(x)}{F_{m,\sigma}(x)} \leqslant \frac{f_{m',\sigma}(x)}{f_{m,\sigma}(x)}. \tag{310}$$

Integrating with respect to x_1 up to x_0 we obtain:

$$f_{m',\sigma}(x_0)\int_{x_0}^{+\infty} f_{m,\sigma}(x_1)dx_1 \leqslant f_{m,\sigma}(x_0)\int_{x_0}^{+\infty} f_{m',\sigma}(x_1)dx_1 \tag{311}$$

$$f_{m',\sigma}(x_0)(1 - F_{m,\sigma}(x_0)) \leqslant f_{m,\sigma}(x_0)(1 - F_{m',\sigma}(x_0)) \tag{312}$$

so for any $x \in \mathbb{R}$ it holds:

$$\frac{f_{m',\sigma}(x)}{f_{m,\sigma}(x)} \leqslant \frac{1 - F_{m',\sigma}(x)}{1 - F_{m,\sigma}(x)}. \tag{313}$$

Putting together the two inequalities yields to:

$$\frac{F_{m',\sigma}(x)}{F_{m,\sigma}(x)} \leqslant \frac{1 - F_{m',\sigma}(x)}{1 - F_{m,\sigma}(x)} \Rightarrow F_{m,\sigma}(x) \geqslant F_{m',\sigma}(x), \quad \forall x \tag{314}$$

Now we tackle Beta distributd rvs. Integrating with respect to x_0 up to x_1 we obtain:

$$f_{\alpha,\beta}(x_1)\int_0^{x_1} f_{\alpha',\beta'}(x_0)dx_0 \leqslant f_{\alpha',\beta'}(x_1)\int_0^{x_1} f_{m,\sigma}(x_0)dx_0 \tag{315}$$

$$f_{\alpha,\beta}(x_1)F_{\alpha',\beta'}(x_1) \leqslant f_{\alpha',\beta'}(x_1)F_{\alpha,\beta}(x_1), \tag{316}$$

so for any $x \in [0, 1]$ it holds:

$$\frac{F_{\alpha',\beta'}(x)}{F_{\alpha,\beta}(x)} \leqslant \frac{f_{\alpha',\beta'}(x)}{f_{\alpha,\beta}(x)}. \tag{317}$$

Integrating with respect to x_1 up to x_0 we obtain:

$$f_{\alpha',\beta'}(x_0)\int_{x_0}^{1} f_{\alpha,\beta}(x_1)dx_1 \leqslant f_{\alpha,\beta}(x_0)\int_{x_0}^{1} f_{\alpha',\beta'}(x_1)dx_1 \tag{318}$$

$$f_{\alpha',\beta'}(x_0)(1 - F_{\alpha,\beta}(x_0)) \leqslant f_{\alpha,\beta}(x_0)(1 - F_{\alpha',\beta'}(x_0)) \tag{319}$$

so for any $x \in [0, 1]$ it holds:

$$\frac{f_{\alpha',\beta'}(x)}{f_{\alpha,\beta}(x)} \leqslant \frac{1 - F_{\alpha',\beta'}(x)}{1 - F_{\alpha,\beta}(x)}. \tag{320}$$

Putting together the two inequalities yields to:

$$\frac{F_{\alpha',\beta'}(x)}{F_{\alpha,\beta}(x)} \leqslant \frac{1 - F_{\alpha',\beta'}(x)}{1 - F_{\alpha,\beta}(x)} \Rightarrow F_{\alpha,\beta}(x) \geqslant F_{\alpha',\beta'}(x), \quad \forall x. \tag{321}$$

Thus obtaining the statements. □

D WHY ENFORCING EXPLORATION IS GENERALLY BENEFICIAL?

Although the optimal value for the exploration parameter for each instance (i.e., $\sigma_{\boldsymbol{\mu}}(T)$) depends on the instance itself[18] and the quantities needed to compute it might not be accessible to the learner, the result from Theorem 6.5 and Corollary 6.4 , makes clear why, even without the knowledge of these quantities, enforcing exploration is beneficial. In fact, although forced exploration may add a source of regret for some instances, it will ensure enough exploration to be able to have tighter guarantees for a wider range of problems. For example, setting an arbitrary forced exploration parameter $\Gamma = \Lambda$ would guarantee enough exploration for the algorithms to properly face all the instances from $\mathcal{M}_\Lambda$. Furthermore, these results also provide a useful criterion for setting the exploration parameter Γ, i.e., whenever the policymaker thinks the class of problems is hard to learn should set Γ high, to ensure that the algorithms can learn a wider range of instances with guarantees on the regret, viceversa whenever the policymaker expects to face simple problems should set Γ low.

---

[18] We highlight that this is common in the forced exploration literature (see, Garivier et al. (2016), Chapter 6 of Lattimore and Szepesvári (2020)), and is not specific of our algorithms.

E NUMERICAL SIMULATIONS, PARAMETERS AND REPRODUCIBILITY DETAILS

E.1 Parameters

The choices of the parameters are those suggested by the authors:

- Rexp3: $V_T = K$ as we've considered bounded rewards within zero and the maximum global variation possible is equal to the number of arms of the bandit; $\gamma = \min\left\{1, \sqrt{\frac{K \log K}{(e-1)\Delta_T}}\right\}, \Delta_T = \left\lceil (K \log K)^{1/3} \left(T/V_T\right)^{2/3} \right\rceil$ Besbes et al. (2014);
- KL-UCB: $c = 3$ as required by the theoretical results on the regret provided by Garivier and Cappé (2011);
- Ser4: according to what suggested by Allesiardo et al. (2017) we selected $\delta = 1/T, \epsilon = \frac{1}{KT}$, and $\phi = \sqrt{\frac{N}{TK \log(KT)}}$;
- SW-UCB: as suggested by Garivier and Cappé (2011) we selected the sliding-window $\tau = 4\sqrt{T \log T}$ and the constant $\xi = 0.6$;
- SW-KL-UCB as suggested by Garivier and Moulines (2011) we selected the sliding-window $\tau = \sigma^{-4/5}$;
- SW-TS: as suggested by Trovò et al. (2020) for the smoothly changing environment we set $\beta = 1/2$ and sliding-window $\tau = T^{1-\beta} = \sqrt{T}$. Even though the paper handle just Bernoulli rewards, we will choose the same order for the window length also for γ-`SWGTS`.
- R-ed-UCB: the window is set as $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$ as suggested by the authors Metelli et al. (2022), $\epsilon \in (0, \frac{1}{2})$, being $N_{i,t-1}$ the numbers of plays of the $i-th$ arm up to time t.

E.2 Environment

To evaluate the algorithms in the rested setting with $K = 15$ arms over a time horizon of $T = 100,000$ rounds. The payoff functions $\mu_i(\cdot)$ have been chosen in these families:

$$F_{\text{exp}} = \left\{ f(n) \mid f(n) = c\left(1 - e^{-an}\right) \right\}, \tag{322}$$

$$F_{\text{poly}} = \left\{ f(n) \mid f(n) = c\left(1 - b\left(n + b^{1/\rho}\right)^{-\rho}\right) \right\}, \tag{323}$$

where $a, c, \rho \in (0, 1]$ and $b \in \mathbb{R}_{\geqslant 0}$ are parameters, whose values have been selected randomly. The complete settings and function selection method, in compliance with what has been presented by Metelli et al. (2022), have been provided in the attached code.

E.3 Experimental Infrastructure

In this section, we provide additional information for the full reproducibility of the experiments provided in the main paper.

The code has been run on an AMD Ryzen 7 4800H CPU with 16 GiB of system memory. The operating system was Windows 11, and the experiments have been run on Python 3.8. The libraries used in the experiments, with the corresponding versions, were:

- `matplotlib==3.3.4`
- `tikzplotlib==0.10.1`
- `numpy==1.20.1`

On this architecture, the average execution time of each algorithm takes an average of $\approx 20 - 40$ sec both in the synthetic environment for a time horizon of $T = 100,000$ and int the IMDB environment for $T = 50000$.

E.4 15-arms Numerical Simulation Results

The results of the numerical simulation presented in Section 8 are reported in Figure 9. In this case we have performed 20 runs, in the semi-transparent areas are reported the standard deviations. The results show how the methods that have been designed for the restless case perform worse than the one we presented in our paper. The only exception is the `Beta-SWTS` that we showed to also have nice theoretical properties in the SRRB setting. Overall, the comparison with such methods do not invalidate the conclusions we drew in the main paper.

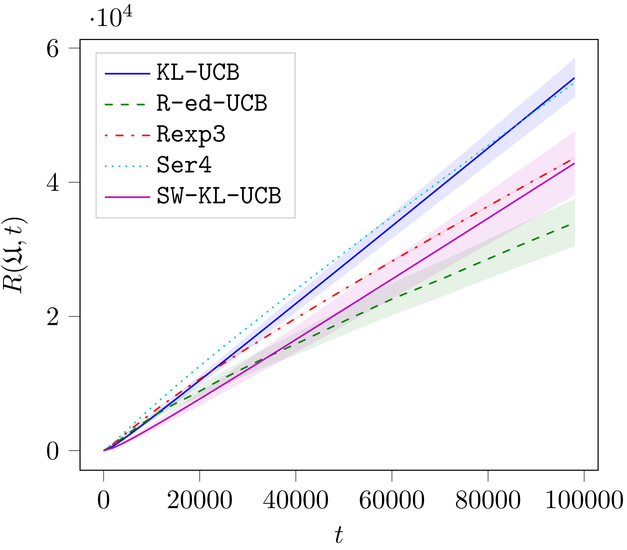


Figure 9: Regret in the 15-arm environment.

E.5 IMDB Experiment

SRRBs are a powerful tool to model a lot of real-world challenges. In particular, we focus on concave rising rested bandits that are suitable to describe *recommendation systems*, in which you should be able to model satiation effects Clerici et al. (2023); Xu et al. (2023). In particular, in this scenario, bandit algorithms serve as meta-algorithms. In fact, we will consider the learning algorithms (such as, logistic regression, neural networks and OGD) as different arms, and the expected reward will be based on the score given by IMDB. We refer the reader to (Metelli et al., 2022, Appendix E.2) for the details of the setting. Specifically, we consider a constant learning rate for Logistic Regression ($\lambda_t = 1$). Moreover, the NNs use as activation functions the rectified linear unit, i.e., $\mathrm{relu}(x) = \max\{0, x\}$, a constant learning rate $\alpha = 0.001$ and the Adam stochastic gradient optimizer for fitting. Two of the chosen nets have only one hidden layer, with 1 and 2 neurons, respectively, the third net has 2 hidden layer, with 2 neurons each, and two nets have 3 layers with 2,2,2 and 1,1,2 neurons, respectively. We refer to a specific NN denoting the cardinalities of the layers, e.g., the one having 2 layer with 2 neurons each is denoted by NN22. We analyzed their global performance on the IMDB dataset by averaging 1,000 independent runs in which each strategy is sequentially fed with all the available 50,000 samples. The learning curves are depicted in Figure 10. The experiments will be organized along this line: first, we run all the algorithms against we are competing comparing them with the results of conventional TS algorihms, with T time horizon set to 50000. The experiment shows that the algorithms devised for the standard bandit are competitive, meaning that the environment is not too challenging (coherent with our analysis, i.e., σ is low). For the second experiment, we add two artificial arms (the details on the artificial arm are provided in the code) to make the environment more challenging for the algorithms. We provide the average cumulative regret and the standard deviation in the semi-transparent areas.

Experiment 1, Figure 11 The experiment shows that, as we have discussed in the paper, conventional TS algorithms can be still effective whenever the complexity induced by the term σ is low enough. In fact `Beta-TS` and standard γ-`GTS` outperform all the algorithms, even those designed explicitly for Rising Rested bandits, i.e., `R-ed-UCB`. Compared to the results presented in the main paper, here we report additional baselines.

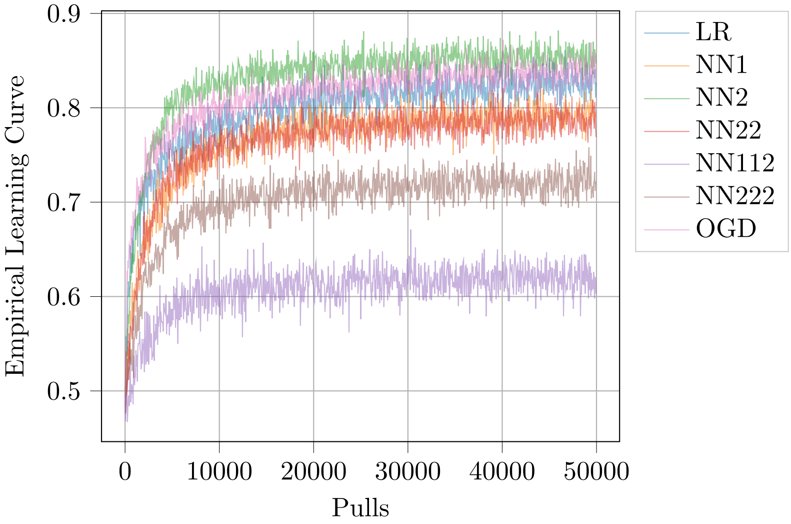


Figure 10: Empirical Learning Curve for each algorithm.

Experiment 2, Figure 12 We add two arms whose evolution challenges the effectiveness of standard bandit algorithms (details are provided in the attached code), i.e., such that a large number of pulls is needed to inferentially estimate the best arm from the other. One of these new two arms is the optimal one. In this experiment we set the forced exploration again equal to 2000, and the sliding window set as suggested in Appendix E. This experiment shows that the conventional algorithms start to fail, as well as the algorithms devised to face restless settings. Conversely, `R-ed-UCB` and the forced exploration TS-like algorithms are able to consistently select the best arm. Differently to what happens for the synthetic setting, `R-ed-UCB` is the best performing algorithm (note that the presence of the additional arms makes the learning problem different compared to the original IMDB setting), however, the TS-like algorithms with forced exploration, shows a more pronounced flattening of the curve, so that for large enough time horizons T we expect them to cath up.

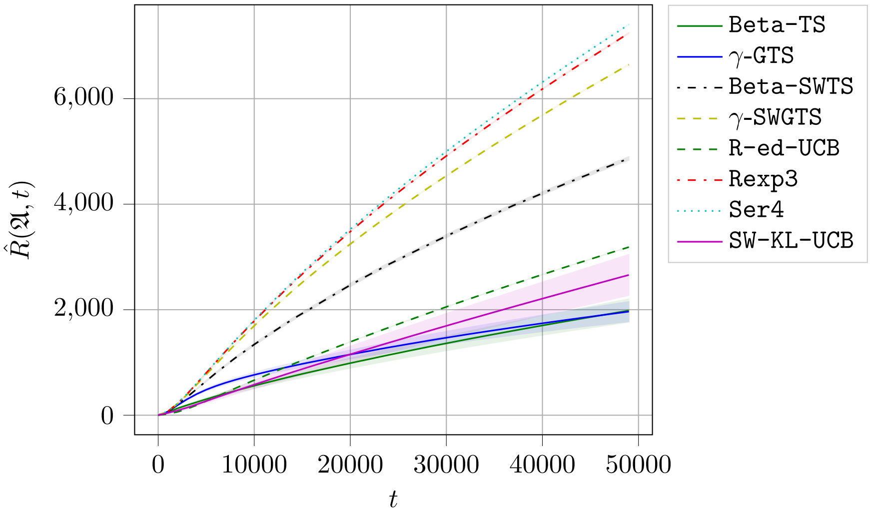


Figure 11: Average Cumulative Regret for the first environment (original IMDB experiment).

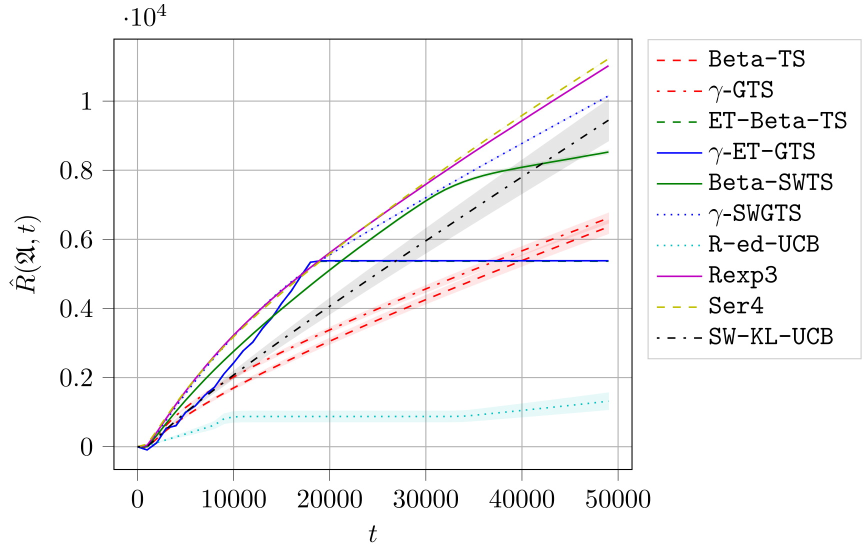


Figure 12: Average Cumulative Regret for the second environment (IMDB experiment with additional arms).

F ABOUT THE COMPUTATIONAL COMPLEXITY

There are three main approaches to handle dynamic environments.

Discounting (Qi et al. (2023); Garivier and Moulines (2011)): giving different weights to the gathered examples, so that some samples can be more important for the learner. This class of methods need to re-compute the average values at every turn for every arm, by updating the weight of all the samples collected during the learning, so that the complexity would scale as $O(t)$. However, the most common and exploited version of the discounted methods can reach $O(1)$ complexity per round per arm in the update phase, as it updates the mean value by a multiplication, so that the overall complexity per round is $O(K)$, being K the number of arms.

Change detection methods (Liu et al. (2018); Besson and Kaufmann (2019)): that at every turn t run a ratio test routine to infer if the distribution is changed, operation that is computationally expensive as it is requested most of the times to compute threshold values that need to be evaluated for every possible subset of samples gathered within $[s, t]$ with $s \in [1, t]$ (see for example Definition 1 of Besson and Kaufmann (2019)), so that a time complexity of $O(t)$ per round is unavoidable.

Sliding Window approaches (Trovò et al. (2020)): are those approaches in which we need to remove the effect of the oldest observation. This requires mantaining a deque to store rewards per arm, leading to efficient removal in $O(1)$ per arm, and therefore to an overall time complexity per round of order $O(K)$, as the sampling itself for the Thompson sample is also $O(1)$ per arm and finding the maximum is $O(K)$. We highlight that is the same time complexity per round of the standard Thompson Sampling used for stationary setting.

Thus, Sliding-Window approaches for Thompson Sampling are one of the most competitive approaches in literature even for what regard time complexity.